EE 274 Lecture 3

Oct. 4, 2023

- Kraft's inequality
- Entropy
- Fundamental limit on lossless compression

Announcements

- IT Forum this Friday
  - 10/6/23, 2pm, Packard 202
  - Jouni Sirén, UCSC on genomic compression
- Code snippets on website
- SCL tutorial
- HW 1

→ Quick Tour

Why SCL?

- Efficient implementations often hard for a beginner to understand or modify
- Implementations of many basic algorithms hard to find
- Intuitively understanding the algorithm ≠ being able to implement it in practice

Why SCL?

- Provide research implementation of common data compression algorithms
- Provide convenient framework to quickly modify existing compression algorithm and to aid research in the area
- To ourselves understand these algorithms better ☺

Quiz

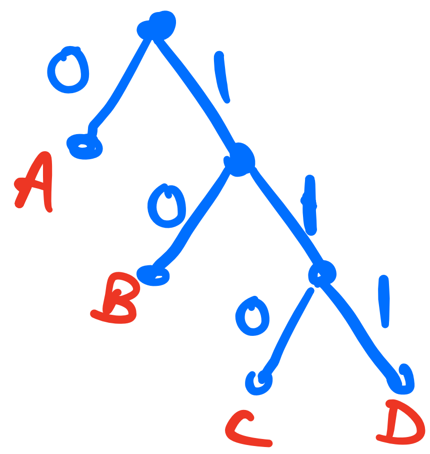
We'll come back in a bit...

Last Time

- Prefix Codes

- Binary Tree Representation
- Good code: $\ell(x) \approx \log_2 \frac{1}{P(x)}$
- Shannon code:

$$\ell(x) = \left\lceil \log_2 \frac{1}{P(x)} \right\rceil$$



Kraft's inequality

Codewords on leaves of binary tree

$\Updownarrow$

Prefix Code

$\Updownarrow$

Lengths satisfy Kraft's inequality

Kraft's inequality

Given a prefix code with codeword lengths $l_1, l_2, \dots, l_k$:

$$\sum_{i=1}^{k} 2^{-l_i} \leq 1$$

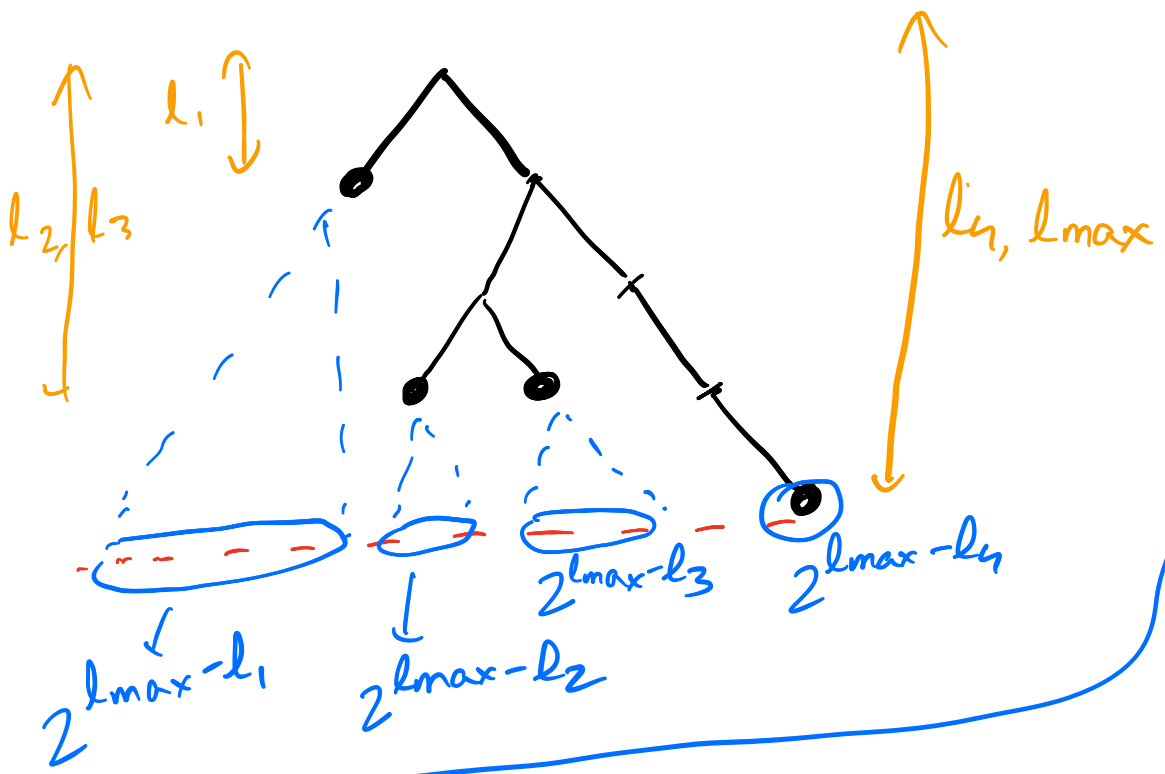
Conversely: Given integers $l_1, \dots l_k \geq 1$ satisfying $\sum_{i=1}^{k} 2^{-l_i} \leq 1$ there exists a prefix code with lengths $l_1, \dots, l_k$

Proof of converse:

If $\sum_{i=1}^{k} 2^{-l_i} \leq 1$, you can use the same code construction as Shannon codes (check the proof!) to get a prefix code with lengths $l_1, \dots, l_k$

Proof of forward part (sketch)

Let $l_{max} = \max\{l_1, \dots, l_k\}$



$2^{l_{max}-l_i}$ = descendants of i^{th} codeword at depth l_{max} ①

Descendants of i & j do NOT overlap (prefix-free) ②

① + ② $\Rightarrow$

$$\sum_{i=1}^{k} 2^{l_{max}-l_i} \leq 2^{l_{max}}$$

$\boxed{\text{sum of sizes of disjoint subsets} \leq \text{total size of set}}$

∎

Quiz

Q1: $\mathcal{X} = \{A, B, C, D, E\}$

1.1

x	$P(x)$	$l(x) = \lceil \log_2 \frac{1}{P(x)} \rceil$	$c(x)$
A	0.25	2	00
B	0.25	2	01
C	0.25	2	10
D	0.13	3	110
E	0.12	4	1110

1.2

$$\mathbb{E}[l(x)] = 0.25 \times 2 + 0.25 \times 2 + 0.25 \times 2 + 0.13 \times 3 + 0.12 \times 4$$

$$= 2.37 \text{ bits/symbol}$$

Quiz

x	$P(x)$	$l(x) = \lceil \log_2 \frac{1}{P(x)} \rceil$
A	0.25	2
B	0.25	2
C	0.25	2
D	0.13	3
E	0.12	4

Kraft sum: $\sum_{i=1}^{5} 2^{-l_i} = 3 \times \frac{1}{2^2} + \frac{1}{2^3} + \frac{1}{2^4}$

$$= \frac{3}{4} + \frac{1}{8} + \frac{1}{16}$$

$$= \frac{15}{16} < 1$$

Quiz

1.3

x	$P(x)$	$\ell(x) = \lceil \log_2 \frac{1}{P(x)} \rceil$	Better Code
A	0.25	2	00
B	0.25	2	01
C	0.25	2	10
D	0.13	3	111
E	0.12	$\cancel{4}$ 3	110

$$E[\ell(x)] = 2.25 < 2.37$$

$$\text{Kraft sum} = 3 \times \frac{1}{2^2} + 2 \times \frac{1}{2^3}$$

$$= \frac{3}{4} + \frac{2}{8} = 1$$

Equality

Quiz

2. Prefix free?

	2.1	2.2	2.3
A	00	001	00
B	01	011	~~00~~ 01
C	10	100	10
D	11	111	111
E	110	110	110
	NO	YES	NO

Kraft sum $= \frac{1}{2^2} \times 4 + \frac{1}{2^3} = 1 + \frac{1}{8} > 1$

$\frac{5}{8} \leq 1$

$\frac{3}{4} + \frac{2}{8} = 1$

Moving on...

- Define information-theoretic quantities
- Precisely characterize the "best" prefix code
- Justify thumb rule: $\ell(x) \approx \log_2 \frac{1}{P(x)}$

Entropy

Let $\mathcal{X} = \{1, \ldots, k\}$, $p_i = P(X=i)$

Entropy $H(X) = \sum_{i=1}^{k} p_i \log_2 \frac{1}{p_i}$ bits

Notes:

1. $H(X)$ is really $H(P)$
2. $H(X) = \mathbb{E}\left[\log_2 \frac{1}{P(X)}\right] \rightarrow \sum p_i \log_2 \frac{1}{P(X=i)}$
3. $p=0$: DEFINE $\Rightarrow p \log_2 \frac{1}{p} = 0$

Examples:

Uniform distribution:

$$X \sim \mathrm{Unif}\{1, \ldots, k\}$$

$$p(x=i) = \frac{1}{k} \quad \forall i$$

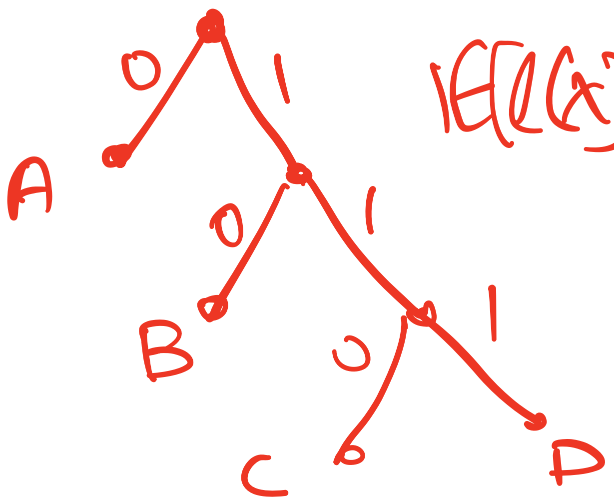
$$H(X) = \sum_{i=1}^{k} \frac{1}{k} \log_2 \frac{1}{1/k} = \log_2 k$$

$$X \sim \mathrm{Ber}(1/2) = \mathrm{Unif}\{0, 1\}$$

$$H(X) = \log_2 2 = 1 \text{ bit}$$

Examples

x	$P(x)$
A	$1/2$
B	$1/4$
C	$1/8$
D	$1/8$



$\mathbb{E}[\ell(x)]$

$$\frac{1}{2}\log_2 \frac{1}{1/2} + \frac{1}{4}\log_2 \frac{1}{1/4} + \frac{1}{8}\log_2 \frac{1}{1/8} \times 2$$

$$= \frac{1}{2}\times 1 + \frac{1}{4}\times 2 + \frac{1}{8}\times 3 + \frac{1}{8}\times 3$$

$$= \frac{1}{2} + \frac{1}{2} + \frac{\cancel{6}\,3}{\cancel{8}\,4} = \frac{7}{4}$$

Properties of entropy

Let $|\mathcal{X}| = k$

$$0 \leq H(X) \leq \log_2 k$$

when

X is deterministic

$1 \times \log_2 \frac{1}{1} = 0$

when

X is uniformly distributed

$k \times \frac{1}{k} \times \log_2 \frac{1}{1/k} = \log_2 k$

What does entropy mean?

- Uncertainty / Randomness
- Amount of information contained
- How much you can compress
- Sampling from distributions &
  Answer to many puzzles!

Joint entropy of independent r.v.s

Just think of (X_1, X_2) as a random variable

$$\underbrace{H(X_1, X_2)}_{Y} = \mathbb{E}\left[\log_2 \frac{1}{P(\underbrace{X_1, X_2}_{Y})}\right]$$

$$= \mathbb{E}\left[\log_2 \frac{1}{P(X_1)\,P(X_2)}\right] \quad (?)$$

$$= \mathbb{E}\log_2 \frac{1}{P(X_1)} + \mathbb{E}\log_2 \frac{1}{P(X_2)}$$

$$= H(X_1) + H(X_2)$$

Joint entropy of independent r.v.s

X_1, X_2 independent $\Rightarrow H(X_1, X_2) = H(X_1) + H(X_2)$

For iid (independent and identically distributed)

$X^n = (X_1, X_2, \ldots X_n) \sim X$

distributed as

$$H(X^n) = \sum_{i=1}^{n} H(X_i) = n H(X)$$

For non-iid, we'll come back later

Relative entropy / KL-divergence

Let $p = (p_1, p_2, \ldots, p_k)$

& $q = (q_1, q_2, \ldots, q_k)$

be probability distributions on

$\mathcal{X} = \{1, \ldots, k\}$

Then

$$D_{KL}(p \| q) = \sum_{i=1}^{k} p_i \log_2 \frac{p_i}{q_i}$$

Examples in Quiz

Relative entropy / KL-divergence

Property: $D_{KL}(p\|q) \geq 0$

with equality iff (if & only if)

$p = q$

Proof: Lecture notes (convexity)

Relative entropy / KL-divergence

- Measure of distance between probability distributions
- Not symmetric! $D(p\|q) \neq D(q\|p)$ in general
- Comes up in ML/generative models loss functions
- Comes up in compression

Main Result of Lossless Compression

1. For every prefix code

$$\mathbb{E}[\ell(X)] \geq H(X)$$

"Lower bound"/ "Converse"

2. Can achieve $\mathbb{E}[\ell(X)] \approx H(X)$

can get arbitrarily close

with prefix codes

"Achievability"

Main Result of Lossless Compression

- Entropy is the fundamental limit of lossless compression.
- Same result applies to class of uniquely decodable codes [HW]
- Next several lectures: algorithms to achieve entropy efficiently
- All this assumes i.i.d. We'll study more general distributions in Lec. 8-10.

Proof of converse

"can't do better than entropy"

Let $X \sim p$ & prefix code with lengths $(l_1, l_2, \dots, l_k)$ $\longrightarrow$ To show $\sum_{i=1}^{k} p_i l_i \geq H(p)$

Let $q_i = c2^{-l_i}$ where $c = \frac{1}{\sum 2^{-l_i}} \geq 1$ (?)

Note: $\sum_{i=1}^{k} q_i = c \sum_{i=1}^{k} 2^{-l_i} = 1$

So $(q_1, \dots q_k)$ is also a distribution

Proof of converse

Now $D_{KL}(p\|q) \geq 0$ (?)

So $\sum_{i=1}^{k} p_i \log_2 \frac{p_i}{q_i} \geq 0$

split the log

$-H(x)$

$\Rightarrow \sum_{i=1}^{k} p_i \log_2 p_i + \sum_{i=1}^{k} p_i \log_2 \frac{1}{q_i} \geq 0$

move $H(x)$ to RHS

$\Rightarrow \sum_{i=1}^{k} p_i \log_2 \frac{1}{c2^{-l_i}} \geq H(x)$

[$q_i := c2^{-l_i}$
Def^n of $H(x)$

$\Rightarrow \mathbb{E}(l(x)) = \sum_{i=1}^{k} p_i l_i \geq \sum_{i=1}^{k} p_i l_i - \sum_{i=1}^{k} p_i \log_2 c \geq H(x)$

(?)

Proof of converse

$$\Rightarrow \mathbb{E}[\ell(x)] = \sum_{i=1}^{k} p_i l_i \geq \sum_{i=1}^{k} p_i l_i - \sum_{i=1}^{k} p_i \log_2 c \geq H(X)$$

($c \geq 1$ by Kraft's inequality)

Thus, $\mathbb{E}[\ell(x)] \geq H(X)$.

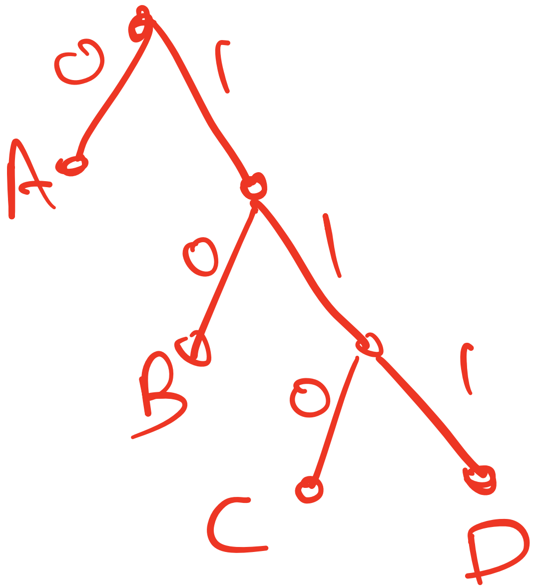
Equality when $c = 1$, $p_i = q_i$

[Recall $D(p\|q) = 0$ iff $p = q$]

$$\Rightarrow p_i = 2^{-l_i} \text{ or } \underbrace{l_i = \log_2 1/p}_{\text{Thumb rule!}}$$

What does $c = 1$ signify? Kraft's with equality!

Kraft's $\longrightarrow$ no p!



$$\sum 2^{-l_i} = 1$$

	$P(x)$
A	$1/8$
B	$1/8$
C	$1/4$
D	$1/2$

$$l_i \neq \log_2 1/p_i$$

If $\ell_i = \log_2 1/p_i$

$$\sum_{i=1}^{k} 2^{-\ell_i} = \sum_{i=1}^{k} 2^{-\log_2 1/p_i}$$

$$= \sum_{i=1}^{k} p_i = 1$$

$\Rightarrow$ Kraft's w/ equality.

Achievability

Shannon codes! $\quad l(x) = \left\lceil \log_2 \frac{1}{p(x)} \right\rceil$

$$\mathbb{E}[l(x)] = \sum_{i=1}^{k} p_i l_i = \sum_{i=1}^{k} p_i \left\lceil \log_2 \frac{1}{p_i} \right\rceil$$

Use $x \leq \lceil x \rceil < x+1$

$$\mathbb{E}[l(x)] < \sum_{i=1}^{k} p_i \left(\log_2 \frac{1}{p(x)} + 1 \right)$$

$$= \sum_{i=1}^{k} p_i \log_2 \frac{1}{p(x)} + \sum_{i=1}^{k} p_i$$

$$= H(x) + 1$$

(?)

$$H(x) \leq \mathbb{E}\, l(x) < H(x) + 1$$

Achievability

Shannon codes! $\quad l(x) = \left\lceil \log_2 \frac{1}{p(x)} \right\rceil$

$$\mathbb{E}[l(x)] < H(x) + 1$$

Shannon code is within 1 bit of entropy!!

Special case: Dyadic distributions
(probabilities powers of 2
like $\frac{1}{2}$, $\frac{1}{4}$ etc.)

$$\mathbb{E}(l(x)) = H(x)$$

Achievability

Shannon codes: $\mathbb{E}[\ell(x)] < H(x)+1$

Not quite "arbitrarily close".

Shannon

A	00
B	01
C	10

Example: Unif({1,2,3})

$H(x) = \log_2 3 = 1.58..$ bits

Shannon code $\mathbb{E}\ell(x) = 2$ bit

$\lceil \log_2 3 \rceil = 2$

Bound $H(x)+1 = 2.58 = H(x)+1$

Overhead $\approx 25\%$.

x	$P(x)$	$\lceil \log_2 1/P(x) \rceil$	$C(x)$	$C^*(x)$
A	$1/3$	2	00	0
B	$1/3$	2	01	01
C	$1/3$	2	10	1

$H(X) = 1.58$ bits

$E[l(x)] = 2$

$$\frac{2}{3} + \frac{2}{3} + \frac{1}{3} = 1.67 \text{ bits}$$

Achievability - not quite there yet

1. Shannon code often suboptimal

   → Huffman code !

2. Even optimal code sometimes far from entropy.

   Block codes

   → Arithmetic codes

   → ANS codes

Block coding (if time permits)

Code in blocks of n symbols:

$$H(X^n) = nH(X)$$

$$H(X^n) \leq E[\ell(X^n)] < H(X^n) + 1$$

$$nH(X) \leq E\,\ell(X^n) < nH(X) + 1$$

$$H(X) \leq \frac{E\,\ell(X^n)}{n} < H(X) + \frac{1}{n}$$

Within $1/n$ of entropy!

Thank You!