PAGE
DATE

Data compression

Def
- $\mathcal{X}$: raw information? data?
- source code or code $C: \mathcal{X} \to \{0,1\}^*$ → only consider binary code.
- $C(x)$: codeword of x, $l(x)$: length of $C(x)$.

~~Example~~
- ~~Assume~~ $\mathcal{X} = \{$ ㄱ, ㄴ, ㄷ, ㄹ $\}$

Classes of codes

① $C(ㄱ)=0$ $C(ㄴ)=0$ $C(ㄷ)=0$ $C(ㄹ)=0$ → Singular

② $C(ㄱ)=0$ $C(ㄴ)=010$ $C(ㄷ)=01$ $C(ㄹ)=10$ → Nonsingular but not uniquely decodable. Consider 010 ⇝ 'ㄴ' or 'ㄱㄹ' or 'ㄷㄱ'.

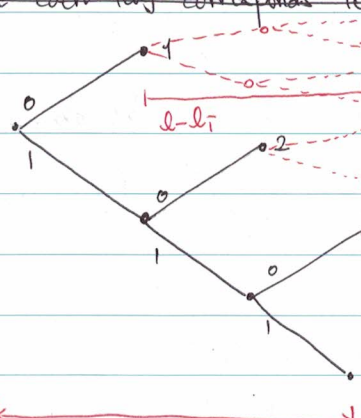
③ $C(ㄱ)=10$ $C(ㄴ)=00$ $C(ㄷ)=11$ $C(ㄹ)=$ ~~110~~
→ Nonsingular, uniquely decodable, not instantaneous.
Consider 110 and 1100 → Need to read the whole string.

④ $C(ㄱ)=0$ $C(ㄴ)=10$ $C(ㄷ)=110$ $C(ㄹ)=111$ → Instantaneous.

WLOG, $\mathcal{X} = \{1, 2, \cdots, m\}$ and write $l_i := l(i)$ $\forall i \in \mathcal{X}$.

Thm 5.2.1 (Kraft ineq.) C is instantaneous iff $\sum_{i\in\mathcal{X}} 2^{-l_i} \le 1$.

pf) (⇒) As C is instantaneous, ∃ corresponding complete binary tree ~~where each leaf corresponds to~~ s.t.



draw imaginary full binary tree.

→ (# leaves) $= 2^l$

(# leaves in i's subtree) $= 2^{l-l_i}$ mutually disjoint

$\Rightarrow \sum_{i\in\mathcal{X}} 2^{l-l_i} \le 2^l$

$\Rightarrow \sum_{i\in\mathcal{X}} 2^{-l_i} \le 1$.

MooKeuk

($\Leftarrow$) $l := \max l_i$, we then have

$$\sum_{i \in \mathcal{X}} 2^{l - l_i} \le 2^l.$$

Use the same technique. $\square$

Problem. Given a random variable X on $\mathcal{X}$, find an instantaneous source code C that minimizes the expected code length

$$\mathbb{E}[l(X)] = \sum_{x \in \mathcal{X}} p(x)\, l(x).$$

Optimal code ~~Formulate the following opt problem:~~

$$\min_{l} \sum_{i \in \mathcal{X}} p_i l_i$$

$$\text{s.t.} \sum_{i \in \mathcal{X}} 2^{-l_i} \le 1 \quad \text{(by Kraft ineq)}$$

$$l_i \in \mathbb{Z}_+, \ \forall i \in \mathcal{X}.$$

- Relaxation:

$$l_i \ge 0 \quad \forall i \in \mathcal{X}.$$

- Further simplification: if $l_i < 0$, $2^{-l_i} > 1$

  $\Rightarrow$ hence, we can drop the sign constraints

$$\min_{l} \sum_{i \in \mathcal{X}} p_i l_i \quad \text{s.t.} \sum_{i \in \mathcal{X}} 2^{-l_i} \le 1 \ \ (\lambda)$$

- Want to find a "lower bound" of the above prob.

$$\min_{l} \underbrace{\sum_{i \in \mathcal{X}} p_i l_i + \lambda \Big(\sum_{i \in \mathcal{X}} 2^{-l_i} - 1 \Big)}_{L} \quad \text{s.t. } \lambda \ge 0$$

$\lambda \ge 0$; ≤ 0 for feas l

- Fix $\lambda \ge 0$, the obj func is differentiable "convex" while $l \in \mathbb{R}^m$. The min is attained when the gradient $= 0$.

$\forall i \in \mathcal{X} \quad \frac{\partial L}{\partial l_i} = p_i - \lambda \cdot 2^{-l_i} \ln 2 = 0. \Rightarrow 2^{-l_i} = \frac{p_i}{\lambda \ln 2}$

- As $\sum_{i \in \mathcal{X}} 2^{-l_i} \le 1$, (and for some reason), $= 1$. ← plug in.

$\Rightarrow \sum_{i \in \mathcal{X}} \frac{p_i}{\lambda \ln 2} = 1 \Rightarrow \lambda = \frac{1}{\ln 2}$. → plug back

- We thus have $\forall i \in \mathcal{X}, \quad l_i = \log \frac{1}{p_i}$. ~~optimal code.~~

~~If $\lceil \log \frac{1}{p_i} \rceil$ are all integral, optimal code! What if not?~~

~~Shannon code~~ ~~$l_i \leftarrow \lceil \log \frac{1}{p_i} \rceil$.~~

~~Relaxation bd~~

Thm ~~For~~ For any instantaneous code C for X, $\mathbb{E}[l(X)] \ge H(X)$.

pf) $l_i \leftarrow \log \frac{1}{p_i}$ is an opt soln to the relaxation. whose obj val is $\sum_{i \in \mathcal{X}} p_i \log \frac{1}{p_i} = H(X)$. ∎

If $\lceil \log \frac{1}{p_i} \rceil$ are all integers, definitely opt! What if not?

Shannon code $l_i \leftarrow \lceil \log \frac{1}{p_i} \rceil$.

Thm This code is an instantaneous code & $\mathbb{E}[l(X)] < H(X) + 1$.

pf) It indeed satisfies Kraft ineq since

$$\sum_{i \in \mathcal{X}} 2^{-l_i} = \sum_{i \in \mathcal{X}} 2^{-\lceil \log \frac{1}{p_i} \rceil} \le \sum_{i \in \mathcal{X}} 2^{-\log \frac{1}{p_i}} = \sum_{i \in \mathcal{X}} p_i = 1.$$

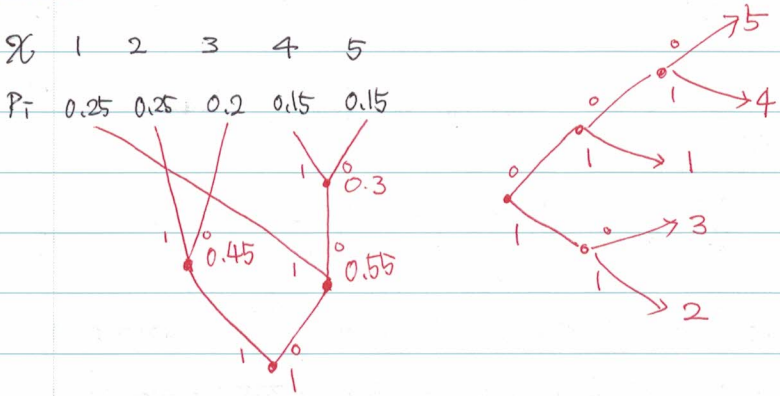
As $l_i < \log \frac{1}{p_i} + 1$,

$$\sum_{i \in \mathcal{X}} p_i l_i < \sum_{i \in \mathcal{X}} p_i \left(\log \frac{1}{p_i} + 1 \right) = H(X) + 1.$$ ∎

Suboptimal Consider $\mathcal{X} = \{1, 2\}$ w/ $p_1 = 1 - \varepsilon$ & $p_2 = \varepsilon$.

Then $l_1 \leftarrow 1$ & $l_2 \leftarrow$ very big. while $C(1) = 0$ $C(2) = 1$ is sufficient.

Huffman code



This code is an integral opt code. Greedy alg. Pf omitted.

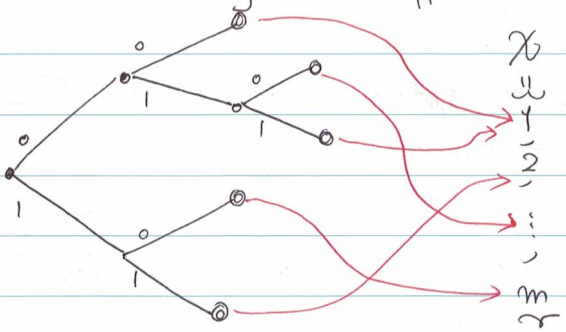
Discrete rand var gen

So far, $X \sim p \longrightarrow$ find C s.t. $\min \mathbb{E}[\ell(X)]$.

Q. # fair coin tosses to generate $X \sim p$?

Thm For any alg to gen $X \overset{\sim p}{,}$ $\mathbb{E}(\#\text{tosses}) \geq H(X)$.

pf) The exec of the alg can be mapped to a complete binary tree.



Let $\mathcal{Y} :=$ (leaves of the tree) & consider mapping $\sigma: \mathcal{Y} \to \mathcal{X}$.

$\forall y \in \mathcal{Y}$, p'_y be the prob of y in tree.

Obsv we have

$$\forall x \in \mathcal{X}, \quad \sum_{y \in \sigma^{-1}(x)} p'_y = p_x.$$

binary tree. & prob.

$$\mathbb{E}[\#\text{tosses}] = \sum_{y \in \mathcal{Y}} p'_y \cdot \ell_y = \sum_{y \in \mathcal{Y}} p'_y \cdot \lg \frac{1}{p'_y}.$$

$$= \sum_{x \in \mathcal{X}} \left(\sum_{y \in \sigma^{-1}(x)} p'_y \lg \frac{1}{p'_y} \right)$$ since $p'_y \le p_x$

$$\ge \sum_{x \in \mathcal{X}} \lg \frac{1}{p_x} \left(\sum_{y \in \sigma^{-1}(x)} p'_y \right) = H(X). \quad \square$$

- If $\{p_x\}$ are all power of two, easy to find an opt alg. What if not? Split into powers of two! e.g.,

$$\frac{7}{8} = \frac{1}{2} + \frac{1}{4} + \frac{1}{8}, \quad \frac{1}{3} = \frac{\frac{1}{4}}{1-\frac{1}{4}} = \frac{1}{4} + \frac{1}{4^2} + \frac{1}{4^3} + \cdots$$

& treat each as distinct outcomes of $\mathcal{Y}$.

Thm For this alg, $\mathbb{E}[\#\text{tosses}] < H(X) + 4$.

pf) $$\mathbb{E}[\#\text{tosses}] = \sum_{y \in \mathcal{Y}} p'_y \cdot \lg \frac{1}{p'_y}$$

$$= \sum_{x \in \mathcal{X}} \left(\sum_{y \in \sigma^{-1}(x)} p'_y \cdot \lg \frac{y p_x}{p'_y} \cdot \frac{1}{p_x} \right)$$

$$= \underbrace{\sum_{x \in \mathcal{X}} \left(\sum_{y \in \sigma^{-1}(x)} p'_y \right) \cdot \lg \frac{1}{p_x}}_{= H(X)} + \underbrace{\sum_{x \in \mathcal{X}} \left(\sum_{y \in \sigma^{-1}(x)} p'_y \cdot \lg \frac{p_x}{p'_y} \right)}_{\le 4?}$$

Suffices to show that $\forall x \in \mathcal{X}$

$$\sum_{y \in \sigma^{-1}(x)} p'_y \cdot \lg \frac{p_x}{p'_y} \le 4 p_x.$$

Spse $\frac{1}{2^{\ell+1}} \le p_x < \frac{1}{2^\ell}$ for some ℓ. Then, a subset of $\{\frac{1}{2^{\ell+1}}, \frac{1}{2^{\ell+2}}, \cdots\}$ constitute $\{p'_y\}$'s; hence,

$$(\text{LHS}) \le \sum_{k=1}^{\infty} \frac{1}{2^{\ell+k}} \cdot \lg \frac{p_x}{\frac{1}{2^{\ell+k}}} < \sum_{k=1}^{\infty} \frac{k}{2^{\ell+k}}$$

PAGE
DATE

As $\sum_{k=1}^{\infty} \frac{k}{2^k} = 2$, we have

$(LHS) < \frac{2}{2^\ell} \leq 4P_\ell$ since $P_\ell \geq \frac{1}{2^{\ell+1}}$. 囧

In textbook, $\mathbb{E}[\#\text{tosses}] \leq H(X)+2$. given.