Information Retrieval

Chandran Goodchild and Abderrahmen Rakez

November 21, 2017

1 Shannon's source coding theorem

$\sum_i 2^{-L_i} \leq 1$, where L_i is the length of the encoding for integer i.
$\Rightarrow$ Show that $E(L_X) \geq H(X)$ when $X \in 1, ..., m$:
Minimize $L_X = \sum_i p_i \cdot L_i$ subject to Kraft's inequality $\sum_i 2^{-L_i} \leq 1$

$$f(L_i, p_i) = \sum_i p_i \cdot L_i \tag{1}$$

$$g(L_i, p_i) = \sum_i 2^{-L_i} - 1 \leq 0 \tag{2}$$

$$p_i = 2^{-L_i} \Rightarrow \mathcal{L} = f - \lambda g = \sum_i p_i \cdot L_i - \lambda \cdot \sum_i p_i - 1 \tag{3}$$

Partial derivatives:

$$\frac{\partial \mathcal{L}(L_i, p_i, \lambda)}{\partial L_i} = \sum_i p_i = 0 \Rightarrow p_i = 0 \tag{4}$$

$$\frac{\partial \mathcal{L}(L_i, p_i, \lambda)}{\partial p_i} = \sum_i L_i - \lambda \sum_i 1 = 0 \Rightarrow L_i = \lambda \tag{5}$$

Rearranging $\sum_i 2^{-L_i} - 1 = \sum_i p_i - 1 \leq 0$

$$\sum_i 2^{-L_i} = \sum_i p_i \leq 1 \tag{6}$$

$$i \cdot 2^{-L_i} = i \cdot p_i \leq 1 \tag{7}$$

$$2^{-L_i} = p_i \leq \frac{1}{i} \tag{8}$$

$$2^{L_i} = \frac{1}{p_i} \geq i \tag{9}$$

$$\Rightarrow L_i = \log_2(\frac{1}{p_i}) \geq \log_2(i) \tag{10}$$

Then:

$$L_X = \sum_i p_i \cdot L_i \geq \sum_i p_i \cdot \log_2 \frac{1}{p_i} = H(x) \quad \square \tag{11}$$

$E(L_x) \leq H(x) + 1$ (part 2 of the source coding theorem) was shown in slide 23 of the lecture.

2 Golomb: Entropy-optimal encoding

X is a fixed gap in an inverted list such that

$$Pr(X = i) = p_i = (1-p)^i \cdot \frac{p}{1-p} \;\mid\; p < 1. \tag{12}$$

Show that Golomb encoding with modulus $M = \frac{1}{p}\ln(2)$ is an entropy-optimal encoding for the gaps $\Rightarrow L_i \leq \log_2(\frac{1}{p_i}) + \mathcal{O}(1)$ for all i.

Golomb:

$$x = q \cdot M + r \tag{13}$$
$$q = \frac{x}{M} \tag{14}$$
$$r = x\%M \tag{15}$$
$$\text{Golomb}(x) = [q]_{\text{unary0}} + 1 + [r]_{\text{binary}} \tag{16}$$

When $M = \frac{1}{p}\ln(2)$:

$$q = \frac{x \cdot p}{\ln(2)} \tag{17}$$
$$r = x\%\frac{\ln(2)}{p} \tag{18}$$

The length of the Golomb code is $L_i = q + 1 + \log_2(r)$

$$L_i = \frac{x \cdot p}{\ln(2)} + 1 + \log_2\left(x\%\frac{\ln(2)}{p}\right) \tag{19}$$
$$x = i \text{ and } p < 1 \Rightarrow L_i \leq \log_2\left(x\%\frac{\ln(2)}{p}\right) + \frac{i}{\ln(2)} + 1 \tag{20}$$
$$1 + x \leq e^x \text{ for } x \in \mathcal{R} \Rightarrow L_i \leq \log_2\left((e^x - 1)\,\%\frac{\ln(2)}{p}\right) + \frac{i}{\ln(2)} + 1 \tag{21}$$
$$L_i \leq \log_2\left(e^x\%\frac{\ln(2)}{p}\right) + \frac{i}{\ln(2)} + 1 \tag{22}$$
$$L_i \leq \log_2\left(\frac{\ln(2)}{p}\right) + \frac{i}{\ln(2)} + 1 \tag{23}$$
$$\log(a \cdot b) = \log(a) + \log(b) \Rightarrow L_i \leq \log_2\left(\frac{1}{p}\right) + \log_2(\ln(2)) + \frac{i}{\ln(2)} + 1 \tag{24}$$
$$L_i \leq \log_2\left(\frac{1}{p_i}\right) + \mathcal{O}(i) \tag{25}$$
$$i = \text{constant / scalar} \Rightarrow L_i \leq \log_2\left(\frac{1}{p_i}\right) + \mathcal{O}(1) \;\; \square \tag{26}$$

3 Space Usage of Optimally Gap-Encoded Inverted Index