Chair for Algorithms
and Data Structures
Prof. Dr. Hannah Bast
Claudius Korzen

Information Retrieval
WS 2017/2018

http://ad-wiki.informatik.uni-freiburg.de/teaching

UNI FREIBURG

Exercise Sheet 4

Solution

Exercise 1 (6 points)

This exercise is about proving one direction of Shannon's famous source coding theorem.

In the lecture, we proved that for any prefix-free code it always holds that $\sum_x 2^{-L_x} \leq 1$, where L_x is the length of the encoding for integer x. Use this to show that for a random variable X drawn from $\{1, \ldots, m\}$, the expected code length $E(L_X)$ is always at least the entropy $H(X)$.

Hint: this is best done with Lagrangian optimization, as explained in Lecture 3.

Solution

We have to show that

$$E(L_X) \geq H(X) = -\sum_{i=1}^{m} p_i \cdot \log_2 p_i \quad \text{where } p_i = Pr(X = i)$$

Proof: Build an instance of Lagrange optimization:

$$\mathcal{L}(i, \lambda) = f(i) + \lambda \cdot (k - g(i))$$

with $f(i) = E(L_X) = \sum_{i=1}^{m} p_i \cdot L_i$ and $g(i) = \sum_{i=1}^{m} 2^{-L_i}$ and $k = 1$. We get:

$$\mathcal{L} = \sum_{i=1}^{m} p_i \cdot L_i + \lambda \cdot \left(1 - \sum_{i=1}^{m} 2^{-L_i}\right)$$

Compute the partial derivatives and set them to zero:

(a)

$$\frac{\partial}{\partial L_i}\mathcal{L} = p_i + \lambda \cdot \ln(2) \cdot 2^{-L_i}$$

$$p_i + \lambda \cdot \ln(2) \cdot 2^{-L_i} \stackrel{!}{=} 0 \tag{1}$$

$$\Leftrightarrow 2^{-L_i} = -\frac{p_i}{\lambda \cdot \ln(2)} \tag{2}$$

(b)

$$\frac{\partial}{\partial\lambda}\mathcal{L} = 1 - \sum_{i=1}^{m} 2^{-L_i}$$

$$1 - \sum_{i=1}^{m} 2^{-L_i} \stackrel{!}{=} 0$$

$$\Leftrightarrow \sum_{i=1}^{m} 2^{-L_i} = 1 \tag{3}$$

Plug (2) into (3) in order to compute the value of λ:

$$\sum_{i=1}^{m} -\frac{p_i}{\lambda \cdot \ln(2)} = 1$$

$$\Leftrightarrow -\frac{1}{\lambda \cdot \ln(2)} \cdot \underbrace{\sum_{i=1}^{m} p_i}_{=1} = 1$$

$$\Leftrightarrow -\frac{1}{\lambda \cdot \ln(2)} = 1$$

$$\Leftrightarrow \lambda = -\frac{1}{\ln(2)} \tag{4}$$

Plug (4) into (1) in order to compute the value of L_i:

$$p_i - \frac{\ln 2}{\ln 2} \cdot 2^{-L_i} = 0$$

$$\Leftrightarrow p_i = 2^{-L_i}$$

$$\Leftrightarrow \log_2 p_i = -L_i$$

$$\Leftrightarrow \log_2 \frac{1}{p_i} = L_i$$

Thus,

$$E(L_X) = \sum_{i=1}^{m} p_i \cdot L_i = \sum_{i=1}^{m} p_i \cdot \log_2 \frac{1}{p_i} = -\sum_{x=1}^{m} p_x \cdot \log_2 p_x = H(X)$$

Check if $E(L_X)$ reaches a global minimum for $L_i = \log_2 \frac{1}{p_i}$:

$$\frac{\partial^2}{\partial L_i^2}\mathcal{L} = \frac{1}{2^{L_i}} \cdot \ln(2)$$

which is > 0 (for $\lambda = -\frac{1}{\ln 2}$). Thus,

$$E(L_X) \geq H(X)$$

□

Exercise 2 (8 points)

This exercise is about the optimal encoding for the gaps of the lists from an inverted index.

In the lecture we have shown that it is reasonable to assume that a fixed gap of an inverted list is distributed like a random variable X with X with $\Pr(X = x) = p_x = (1-p)^{x-1} \cdot p$ for some $p < 1$. Show that under this assumption, Golomb encoding with modulus $M = \lceil 1/p \cdot \ln 2 \rceil$ is an entropy-optimal encoding for the gaps, that is, $L_X \leq \log_2(1/p_x) + O(1)$.

Optionally, if you feel up to it, try to prove the stronger $L_X \leq \log_2(1/p_x) + 1$, or at least try to make the additive constant as small as possible. This is not necessary to get full points for this exercise.

Hint: you can use without proof that $1 + x \leq e^x$ for all real numbers x.

Solution

For Golomb encoding, L_X is defined as (see slide 13):

$$\begin{aligned}
L_X &= \left\lfloor \frac{x}{M} \right\rfloor + 1 + \lceil \log_2 M \rceil \\
&= \underbrace{\left\lfloor \frac{x}{\lceil 1/p \cdot \ln 2 \rceil} \right\rfloor}_{\leq \frac{x}{1/p \cdot \ln 2}} + 1 + \underbrace{\left\lceil \log_2 \left\lceil \frac{1}{p} \cdot \ln 2 \right\rceil \right\rceil}_{\leq \log_2(\frac{1}{p} \cdot \ln 2) + 1} \\
&\leq \frac{x}{1/p \cdot \ln 2} + \log_2(\frac{1}{p} \cdot \ln 2) + 2 \\
&= \frac{x \cdot p}{\ln 2} + \log_2 \frac{1}{p} + \underbrace{\log_2 \ln 2}_{<0} + 2 \\
&\leq \frac{x \cdot p}{\ln 2} + \log_2 \frac{1}{p} + 2 \\
&= (x-1) \cdot \frac{p}{\ln 2} + \underbrace{\frac{p}{\ln 2}}_{\leq 2} + \log_2 \frac{1}{p} + 2 \\
&\leq (x-1) \cdot \frac{p}{\ln 2} + \log_2 \frac{1}{p} + 4 \qquad (5)
\end{aligned}$$

From $p_x = (1-p)^{x-1} \cdot p$ we get:

$$\begin{aligned}
p_x &= (1-p)^{x-1} \cdot p \\
\Leftrightarrow \log_2 p_x &= (x-1) \cdot \log_2(1-p) + \log_2 p \\
\Leftrightarrow -\log_2 p_x &= (x-1) \cdot \log_2(\frac{1}{1-p}) + \log_2 \frac{1}{p} \\
\Leftrightarrow \log_2 \frac{1}{p_x} &= (x-1) \cdot \log_2(\frac{1}{1-p}) + \log_2 \frac{1}{p} \qquad (6)
\end{aligned}$$

From (5) and (6) we can conclude that

$$\begin{aligned} L_X \leq \log_2(1/p_x) + O(1) &\Leftrightarrow \frac{p}{\ln 2} \leq \log_2(\frac{1}{1-p}) \\ &\Leftrightarrow \frac{p}{\ln 2} \leq \frac{\ln(\frac{1}{1-p})}{\ln 2} \\ &\Leftrightarrow p \leq \ln(\frac{1}{1-p}) \\ &\Leftrightarrow e^p \leq \frac{1}{1-p} \\ &\Leftrightarrow 1-p \leq e^{-p} \end{aligned}$$

which holds according to the hint. □

Exercise 3 (6 points)

This exercise is about calculating the space usage of an optimally gap-encoded inverted index (doc ids only, no scores) for a document collection with a total of N words from a vocabular of m words.

Let us make the following assumptions. Let $L_1, \ldots, L_m$ denote the m inverted lists in order of descending lengths (longest list first), and note that $\sum_{j=1}^{m} |L_j| = N$. We assume that the list lengths are Zipf-distributed, that is, $|L_j|$ is proportional to $1/j$. We also assume that the gaps in each inverted list are randomly distributed, as explained in the lecture and assumed in Exercise 2, and that we have an optimal code for each L_j (Golomb with the right modulus would be such an optimal code, but that is not important for this exercise). Under these assumptions, the expected code length for a gap from L_j is $\log_2 j + O(1)$ bits. You can use this without proof.

Show that under these assumptions the expected total number of bits required to gap-encode all the inverted lists is $N \cdot (\log_2 m)/2 + O(N)$. That is, the average number of bits per posting is $(\log_2 m)/2 + O(1)$.

Hint: you can use without proof that $\sum_{j=1}^{m} 1/j = \ln m + O(1)$ and $\sum_{j=1}^{m} (\ln j)/j = (\ln^2 m)/2 + O(1)$.

Solution

Since $|L_j|$ is proportional to $1/j$, $|L_j| = A \cdot N \cdot \frac{1}{j}$ and

$$N = \sum_{j=1}^{m} |L_j| = A \cdot N \cdot \sum_{j=1}^{m} \frac{1}{j}$$

for any constant $A > 0$. It follows:

$$\begin{aligned} N &= A \cdot N \cdot \sum_{j=1}^{m} \frac{1}{j} \\ \Leftrightarrow A &= \frac{N}{N \cdot \sum_{j=1}^{m} \frac{1}{j}} \\ \Leftrightarrow A &= \frac{1}{\sum_{j=1}^{m} \frac{1}{j}} \end{aligned}$$

Thus,

$$\sum_{j=1}^{m} |L_j| = \frac{N}{\sum_{j=1}^{m} \frac{1}{j}} \cdot \sum_{j=1}^{m} \frac{1}{j} \tag{7}$$

Let X be the number of bits required to gap encode all the inverted lists $L_1, \ldots, L_m$. Then,

$$\begin{aligned} E(X) &= \sum_{j=1}^{m} |L_j| \cdot (\log_2 j + B) \qquad \text{(with constant } B \geq 0) \\ &= \sum_{j=1}^{m} |L_j| \cdot \log_2 j + B \cdot \underbrace{\sum_{j=1}^{m} |L_j|}_{=N} \\ &= \sum_{j=1}^{m} |L_j| \cdot \log_2 j + O(N) \\ &= \frac{N}{\sum_{j=1}^{m} \frac{1}{j}} \cdot \sum_{j=1}^{m} \frac{\log_2 j}{j} + O(N) \qquad \text{(with (7))} \\ &= \frac{N}{\sum_{j=1}^{m} \frac{1}{j}} \cdot \sum_{j=1}^{m} \frac{\frac{\ln j}{j}}{\ln 2} + O(N) \\ &= \frac{N}{\sum_{j=1}^{m} \frac{1}{j} \cdot \ln 2} \cdot \sum_{j=1}^{m} \frac{\ln j}{j} + O(N) \\ &= \frac{N}{\ln 2} \cdot \frac{\frac{\ln^2 m}{2} + O(1)}{\ln m + O(1)} + O(N) \qquad \text{(with hint)} \\ &= \frac{N}{\ln 2} \cdot \frac{\frac{\ln^2 m}{2}}{\ln m} + O(1) + O(N) \\ &= \frac{N}{\ln 2} \cdot \frac{\ln m}{2} + O(N) \\ &= \frac{N}{2} \cdot \frac{\ln m}{\ln 2} + O(N) \\ &= \frac{N}{2} \cdot \log_2 m + O(N) \end{aligned}$$

□