Exercise Sheet 4 solutions

Przemyslaw Joniak, pj46

November 21, 2017

Exercise 1

Given L_i $(1 < i \leq N)$ prove that $E(L_X) \geq H(X)$. I will prove this using *Lagrange multiplier* technique. Following equations will be useful:

$$E(L_X) = \sum_i p_i L_i \qquad \textit{(Expectation def.)} \tag{1}$$

$$\sum_i 2^{-L_i} \leq 1 \qquad \textit{(Kraft's inequality)} \tag{2}$$

Proof. Let $S = \sum_i 2^{-L_i}$. Now a constraint equation can be written: $\sum_i 2^{-L_i} - S = 0$. Having both function to minimize and the constraint, a *Lagrange function* can be defined as follow:

$$\mathcal{L}(L_1, L_2, ..., L_N, \lambda) = \sum_i p_i L_i + \lambda(\sum_i 2^{-L_i} - S)$$

Taking partial derivatives and comparing to zero:

$$\frac{\partial \mathcal{L}}{\partial L_i} = p_i - \lambda 2^{-L_i} \log 2 = 0$$

we get:

$$2^{-L_i} = \frac{p_i}{\lambda \log 2}$$

Use *Kraft's inequality* and a fact that $\sum_i p_i = 1$:

$$\sum_i 2^{-L_i} = \sum_i \frac{p_i}{\lambda \log 2} = \frac{1}{\lambda \log 2} \leq 1$$

Hence:

$$\begin{aligned} 2^{-L_i} = p_i \frac{1}{\lambda \log 2} &\leq p_i \\ 2^{-L_i} &\leq p_i \\ \frac{1}{p_i} &\leq 2^{L_i} \\ L_i &\geq \log_2 \frac{1}{p_i} = -\log_2 p_i \end{aligned}$$

Thus, we get:

$$\begin{aligned} E(L_X) &= \sum_i p_i L_i \\ &\geq -\sum_i p_i \log_2 p_i \\ &= H(X) \end{aligned}$$

$\square$

Exercise 3

What we know:

$$\sum_{j=1}^{m} \frac{1}{j} = \ln m + O(1) \tag{3}$$

$$\sum_{j=1}^{m} \frac{\ln j}{j} = \frac{\ln^2 m}{2} + O(1) \tag{4}$$

$$|L_1| \geq |L_2| \geq \cdots \geq |L_m| \tag{5}$$

$$|L_j| \sim \frac{1}{j} \qquad \text{(List length is Zipf's distributed)} \tag{6}$$

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

Observation: Since list length is *proportional* to $\frac{1}{j}$, then there must exist a constant c such that:

$$|L_j| = c\frac{1}{j}$$

Note that for $j = 1$ it holds: $|L_1| = c$. Since lists are sorted in descending order, than $c = |L_1|$ Putting it into equation (7):

$$\sum_{j=1}^{m} |L_j| = \sum_{j=1}^{m} \frac{|L_1|}{j} \tag{8}$$

Now we enough knowledge to solve to exercise. We are suppose to calculate expected total number of bits required to gap-encode all the inverted lists. In the other words, we have to multiply expected number of gaps in a list by length of the list, and then sum up results for the all lists. Since expected code length for a gap from L_j is $\log_2 j + O(1)$:

$$
\begin{aligned}
\sum_{j=1}^{m}(\log_2 j + O(1))|L_j| &= \sum_{j=1}^{m}\log_2 j|L_j| + \sum_{j=1}^{m}|L_j| \\
&= |L_1|\sum_{j=1}^{m}\frac{\log_2 j}{j} + N \\
&= |L_1|\frac{1}{\ln 2}\sum_{j=1}^{m}\frac{\ln j}{j} + N \\
&= |L_1|\frac{1}{\ln 2}(\frac{\ln^2 m}{2} + O(1)) + N
\end{aligned}
$$

What have I done so far? First, I applied (7)-th and (8)-th equation. Then I changed logarithm basis, and at the end I used (4). Let's continue calculations. In the first step I'll switch a logarithm basis and then I'll apply equation (3)

$$
\begin{aligned}
|L_1|\frac{1}{\ln 2}(\frac{\ln^2 m}{2} + O(1)) + N &= \frac{|L_1|}{\ln 2}(\frac{\ln 2 \log_2 m \ln m}{2} + O(1)) \\
&= \frac{\log_2 m}{2}|L_1|\ln m + \frac{|L_1|}{\ln 2} + O(1) + N \\
&= \frac{\log_2 m}{2}|L_1|(\sum_{j=1}^{m}\frac{1}{j} - O(1)) + \frac{|L_1|}{\ln 2} + O(1) + N \\
&= \frac{\log_2 m}{2}(\sum_{j=1}^{m}\frac{|L_1|}{j} - |L_1|) + \frac{|L_1|}{\ln 2} + O(1) + N \\
&= \frac{\log_2 m}{2}(N - |L_1|) + \frac{|L_1|}{\ln 2} + O(1) + N \\
&= N\frac{\log_2 m}{2} - |L_1|\frac{\log_2 m}{2} + \frac{|L_1|}{\ln 2} + O(1) + N \\
&= N\frac{\log_2 m}{2} - |L_1|\frac{\log_2 m}{2} + O(N) \\
&\leq N\frac{\log_2 m}{2} + O(N)
\end{aligned}
$$

Exercise 2

First, consider remainder part of coding. Its length is: $\lceil \log_2 M \rceil$

$$
\begin{aligned}
\lceil \log_2 M \rceil &= \lceil \log_2 \lceil \frac{\ln 2}{p_i} \rceil \rceil \\
&\leq \lceil \log_2 (\frac{\ln 2}{p_i} + 1) \rceil \\
&\leq \log_2 (\frac{\ln 2}{p_i} + 1) + 1
\end{aligned}
$$

Now note that, for $x > \ln 2$ the following is always true: $x + 1 \leq 4x$. Let's apply it into logarithm arguments:

$$\begin{aligned}\log_2(\frac{\ln 2}{p_i} + 1) + 1 &\leq \log_2 \frac{\ln 2}{p_i} + 3 \\ &= \log_2 \frac{1}{p_i} + \log_2 \ln 2 + 3\end{aligned}$$

Consider first part of coding, the one written in unary. It's length is $\lfloor \frac{i}{M} \rfloor \leq \lceil \frac{i}{M} \rceil = \lceil \frac{i}{\lceil \frac{\ln 2}{p_i} \rceil} \rceil$ Since $i \in \mathbb{Z}$ and right hand side expression is continuous, monotonically increasing function, we can apply theorem 3.10 from Graham, Knuth and Pathashnik's *Concrete Mathematics* (2nd edition, p.71):

$$\lceil \frac{i}{\lceil \frac{\ln 2}{p_i} \rceil} \rceil = \lceil \frac{ip_i}{\ln 2} \rceil$$

Of course: $\lceil \frac{ip_i}{\ln 2} \rceil \leq \frac{ip_i}{\ln_2} + 1 = \frac{1}{\ln_2} i(1-p)^{i-1} + 1 \leq i(1-p)^{i-1} + 1$.

Calculate L_i:

$$\begin{aligned}L_i &= \lfloor \frac{i}{M} \rfloor + 1 + \lceil \log_2 M \rceil \\ &\leq i(1-p)^{i-1} + 1 + 1 + \log_2 \frac{1}{p_i} + \log_2 \ln 2 + 3 \\ &= i(1-p)^{i-1} + \log_2 \frac{1}{p_i} + O(1)\end{aligned}$$

We want this code to be optimal. It'd be great it following inequality is true:

$$i(1-p)^{i-1} + \log_2 \frac{1}{p_i} + O(1) \leq \log_2 \frac{1}{p_i} + O(1)$$

It is true on condition $i(1-p)^{i-1}$ is $O(1)$. Intuitively it's true: for $i = 1$ we get 1. If i increases, then $(1-p)^{i-1}$ decreases drastically (since $p \in [0, 1]$)