Exercise 6-1  Lossy Counting

Before stream clustering, let’s take a look at a more fundamental task in stream: count the occurrence of objects in a stream and output objects with a count larger or equal to some given threshold: $minSup \times L$, where $L$ is the length of the stream up to now and $minSup$ is the given threshold (minimum support).

Lossy Counting is one of the basic algorithms that solve this problem. Given the windows size as $w = \frac{1}{2^k}$, the lossy counting algorithm works as the follows: cut stream into windows, process one window a time and prune histogram entries with 0 counts at each window boundary. The illustration is given below:

Please prove that

(a) the maximum count error (the maximum difference between the real count and the estimated count) of the lossy counting algorithm is $\epsilon L$
Let \( W \) be the current windows id, then \( \frac{1}{\epsilon} W = L \Rightarrow W = \epsilon L \). The count error of a given object is happened due to the reduction of count at each window boundary. Thus, including the current windows boundary, the maximum count error is \( \Delta = W = \epsilon L \).

(b) the memory consumption, i.e., the number of entries stored in the histogram, is \( O\left(\frac{1}{\epsilon} \log(\epsilon L)\right) \). (Optional)

Let \( W \) be the current window id. For each \( i \in [1, W] \), let \( d_i \) denote the number of entries in the histogram \( H \) which corresponds to window \( W - i + 1 \).

Thus, the item in the stream corresponding to such entry must occur at least \( i \) times in window \( B - i + 1 \) through \( W \); otherwise, it would have been deleted. Since the size of each window is \( w \), we have:

\[
\sum_{i=1}^{j} d_i \leq j w \quad \text{for } j = 1, 2, \ldots, W
\]  

(1)

Now we want to prove:

\[
\sum_{i=1}^{j} d_i \leq \sum_{i=1}^{j} \frac{w}{i}
\]  

(2)

By induction:

- For \( j = 1 \), this is true.
- Assume it is true for \( j = 1, 2, \ldots, p - 1 \), then for \( j = p \), adding the inequality (1) for \( j = p \) to the aggregation of all \( p - 1 \) instances of the inequality (2) gives us:

\[
\sum_{i=1}^{p} i d_i + \sum_{i=1}^{1} d_i + \sum_{i=1}^{2} d_i + \cdots + \sum_{i=1}^{p-1} d_i \leq pw + \sum_{i=1}^{1} \frac{w}{i} + \sum_{i=1}^{2} \frac{w}{i} + \cdots + \sum_{i=1}^{p-1} \frac{w}{i}
\]

The left hand side can be written as:

\[
1 \cdot d_1 + 2 \cdot d_2 + \cdots + p \cdot d_p + (d_1 + d_2) + \cdots + (d_1 + d_2 + \cdots + d_{p-1}) = p \cdot d_1 + p \cdot d_2 + \cdots + p \cdot d_p = p \sum_{i=1}^{p} d_i
\]

Similarly, the right hand side can be written as:

\[
\frac{1 \cdot w}{1} + \frac{2 \cdot w}{2} + \cdots + \frac{p \cdot w}{p} + \left(\frac{w}{1} + \frac{w}{2} + \cdots + \frac{w}{p-1}\right) = \frac{pw}{1} + \frac{pw}{2} + \cdots + \frac{pw}{p} = p \sum_{i=1}^{p} \frac{w}{i}
\]

Then we have:

\[
p \sum_{i=1}^{p} d_i \leq p \sum_{i=1}^{p} \frac{w}{i}
\]

\[
\Rightarrow \sum_{i=1}^{p} d_i \leq \sum_{i=1}^{p} \frac{w}{i}
\]

Thus the memory consumption at window \( W \) is \( |H| = \sum_{i=1}^{W} d_i \leq \sum_{i=1}^{W} \frac{w}{i} = \frac{1}{\epsilon} \log W = \frac{1}{\epsilon} \log \epsilon L \)

(Here the inequality \( \sum_{i=1}^{W} \frac{1}{i} \leq \log W \) is used, as it is the harmonic series.)