Ludwig-Maximilians-Universität München
Institut für Informatik
Prof. Dr. Matthias Renz
Felix Borutta, Evgeniy Faerman, Christian Frey, Klaus Schmid

Munich, 14.12.2015

Big Data Management and Analytics

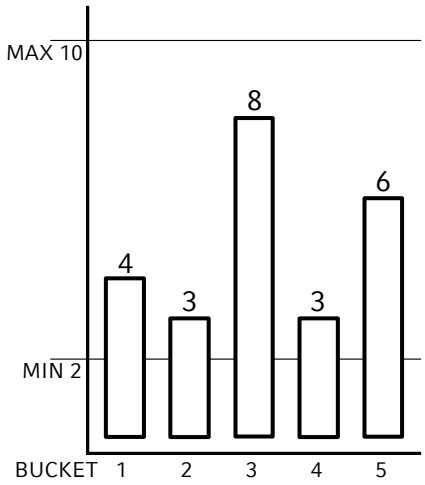
WS 2015/16

Tutorial 6: Stream Algorithms

Assignment 6-1 *K-Buckets* **3 Punkte**

Given the histogram as seen below, execute the K-Buckets Histogram algorithm for inserts and deletes, assuming the following rules:

- The histogram consists of constantly $k = 5$ buckets.
- The upper threshold (MAX) per bucket is 10, the lower threshold (MIN) is 2.
- For split-and-merge operations: a split occurs when the size of a bucket would otherwise **exceed** MAX; a merge occurs between the two consecutive buckets that were not product of the preceding split with the lowest overall sum of sizes.
- For merge-and-split operations: a merge occurs with the neighbour bucket that has the smallest size, when the size of a bucket would otherwise be below MIN.



INSERTING Insert the items of the given sequence into the histogram, until the first overflow occurs. Execute the resulting split-and-merge and move on to the next section (deleting). Each item is denoted as the index of its respective bucket.

$$\text{Sequence} = 3, 1, 3, 5, 2, 3, 4, 1, 5, 3$$

DELETING Starting with the resulting histogram of the insert section, remove the items of the given sequence from the histogram, until the first underflow occurs. Execute the resulting merge-and-split. Each item is denoted as the index of its respective bucket.

$$\text{Sequence} = 1, 3, 4, 5, 4, 3, 2, 5, 1, 2$$

Assignment 6-2 *CUSUM – Change Detection* **3 Punkte**

Given a mean value $\omega = 3$ and a threshold value $\alpha = 8$, execute the Cumulative Sum algorithm for change detection on the following sequence:

$$\text{Sequence} = 2, 3, 7, 4, 0, 2, 5, 6, 8, 7$$

n	x_n - ω	G_n
0		
1		
2		
3		
4		
5		
6		
7		
8		
9		
10		

Assignment 6-3 *Proof by Induction – Lossy Counting Algorithm* **0 Punkte**

For the Frequent Itemset Mining algorithm Lossy Counting, prove the following statement using Induction and the notation from lecture slides:

Whenever an entry (e, f, Δ) gets deleted from the lookup table, the exact frequency $f_e \leq b_{curr}$.

Assignment 6-4 *Exponential Histograms* **0 Punkte**

For the given sequence, construct an Exponential Histogram using a window size $N = 8$ and an error parameter $\epsilon = 1/2$.

Sequence = $\times, \times, \circ, \times, \circ, \circ, \times, \times, \times, \times, \circ, \times, \times, \circ, \times, \times$

Estimate the number of $\times$ within the window at time $t = 13$ and compare it to the actual number.