Big Data Management and Analytics
Assignment 6

Assignment 6-3

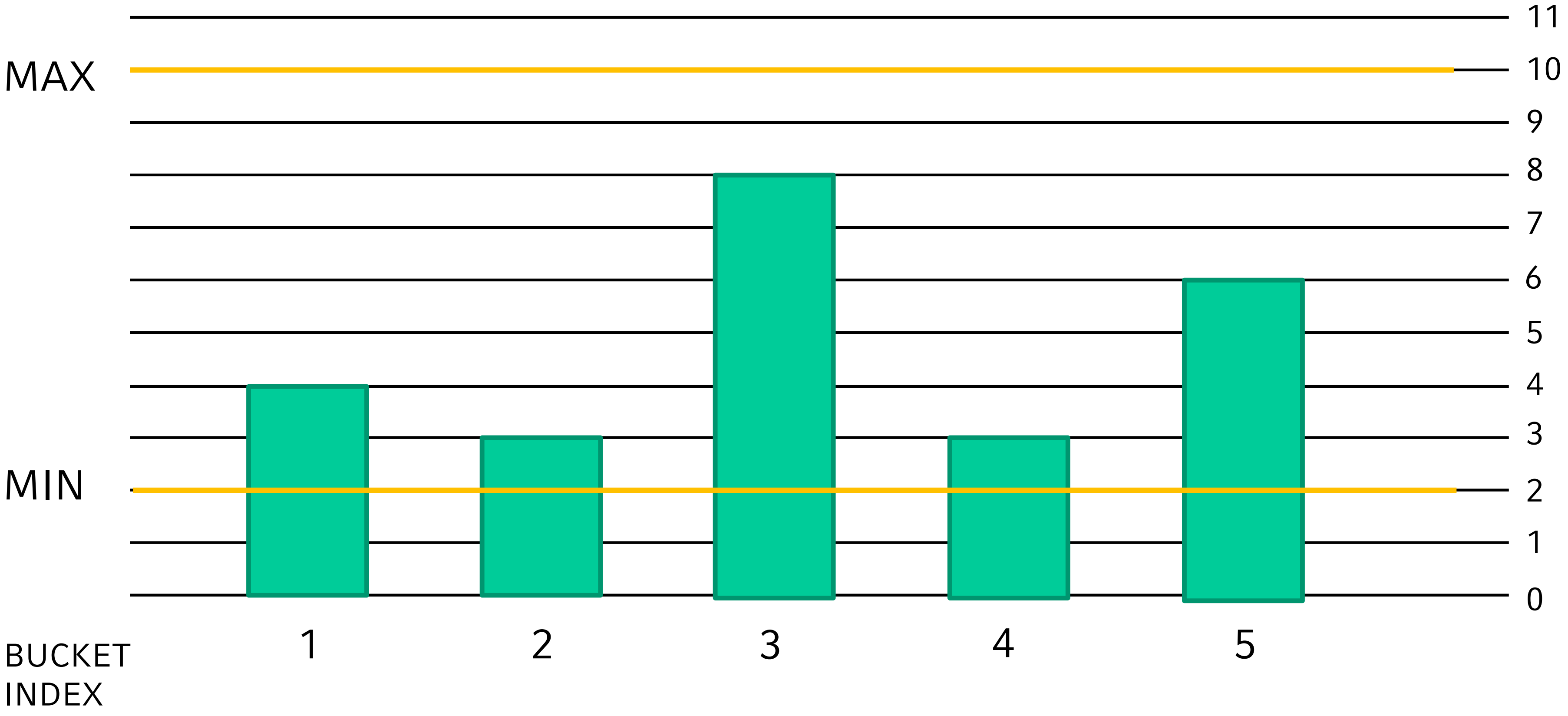
(a) k-Bucket histograms:

- Histogram consists constantly of k=5 buckets
- Upper threshold per bucket MAX = 10
- Lower threshold per bucket MIN = 2

Assignment 6-3

Sequence = 3, 1, 3, 5, 2, 3, 4, 1, 5, 3

Mode: INSERTING

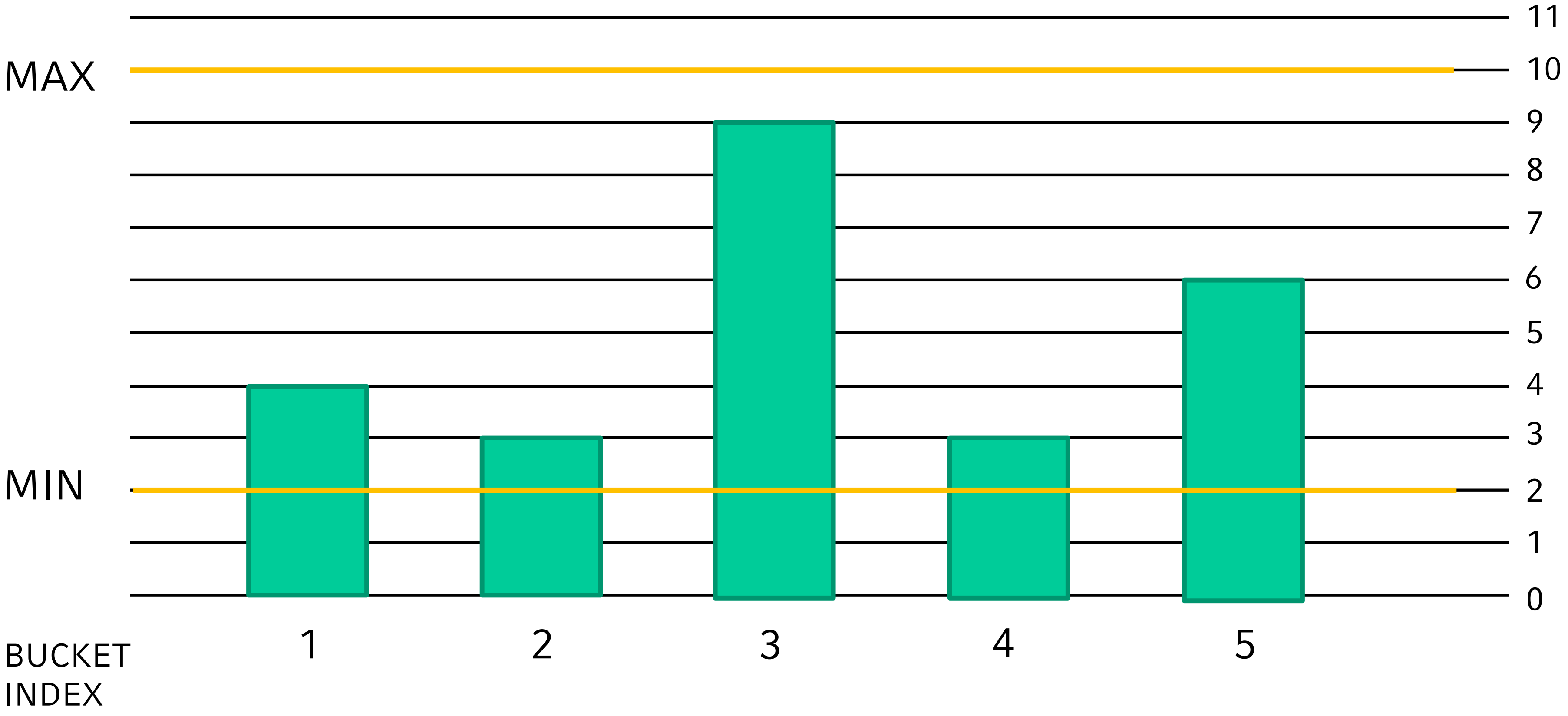


Assignment 6-3

Sequence = 3, 1, 3, 5, 2, 3, 4, 1, 5, 3

Mode: INSERTING

INSERT 3

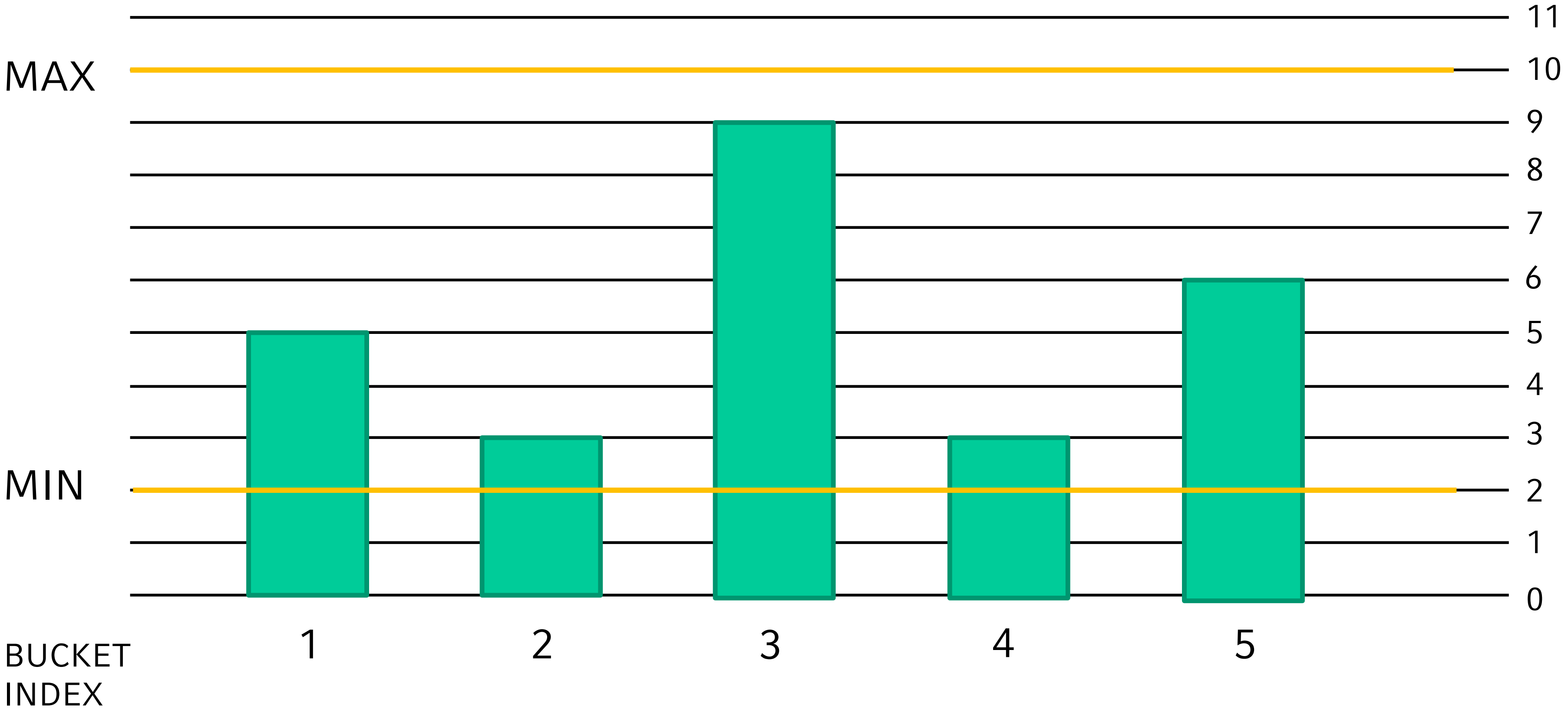


Assignment 6-3

Sequence = 3, 1, 3, 5, 2, 3, 4, 1, 5, 3

Mode: INSERTING

INSERT 1

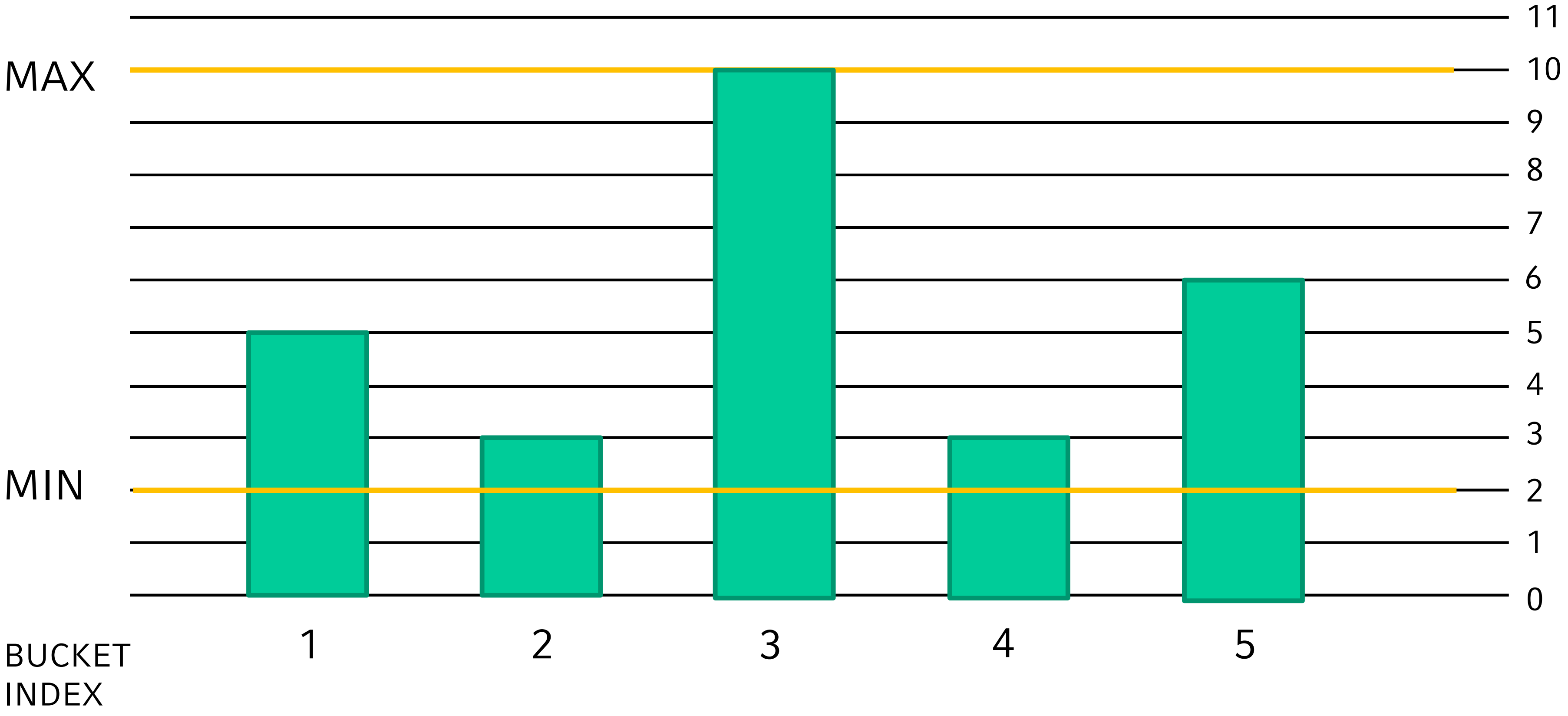


Assignment 6-3

Sequence = 3, 1, 3, 5, 2, 3, 4, 1, 5, 3

Mode: INSERTING

INSERT 3

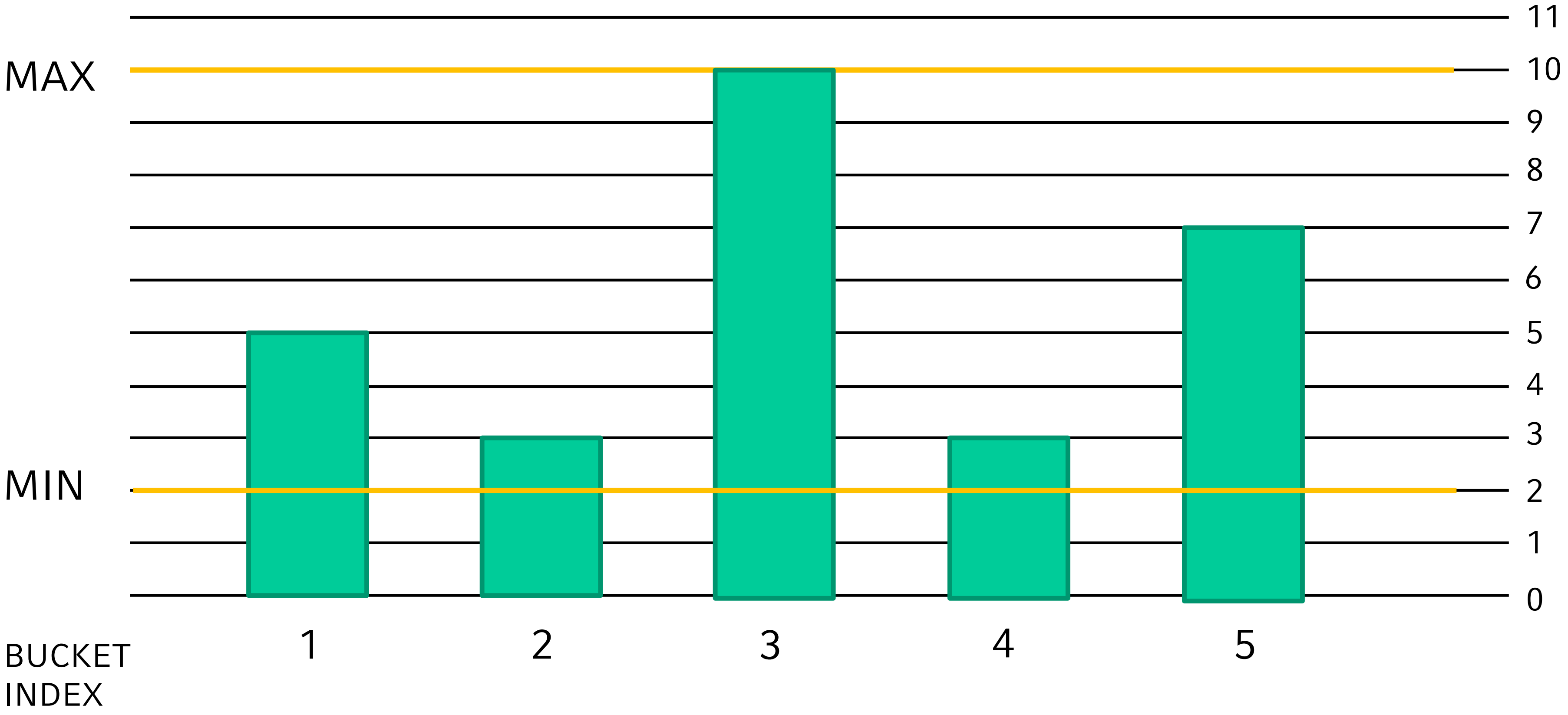


Assignment 6-3

Sequence = 3, 1, 3, 5, 2, 3, 4, 1, 5, 3

Mode: INSERTING

INSERT 5

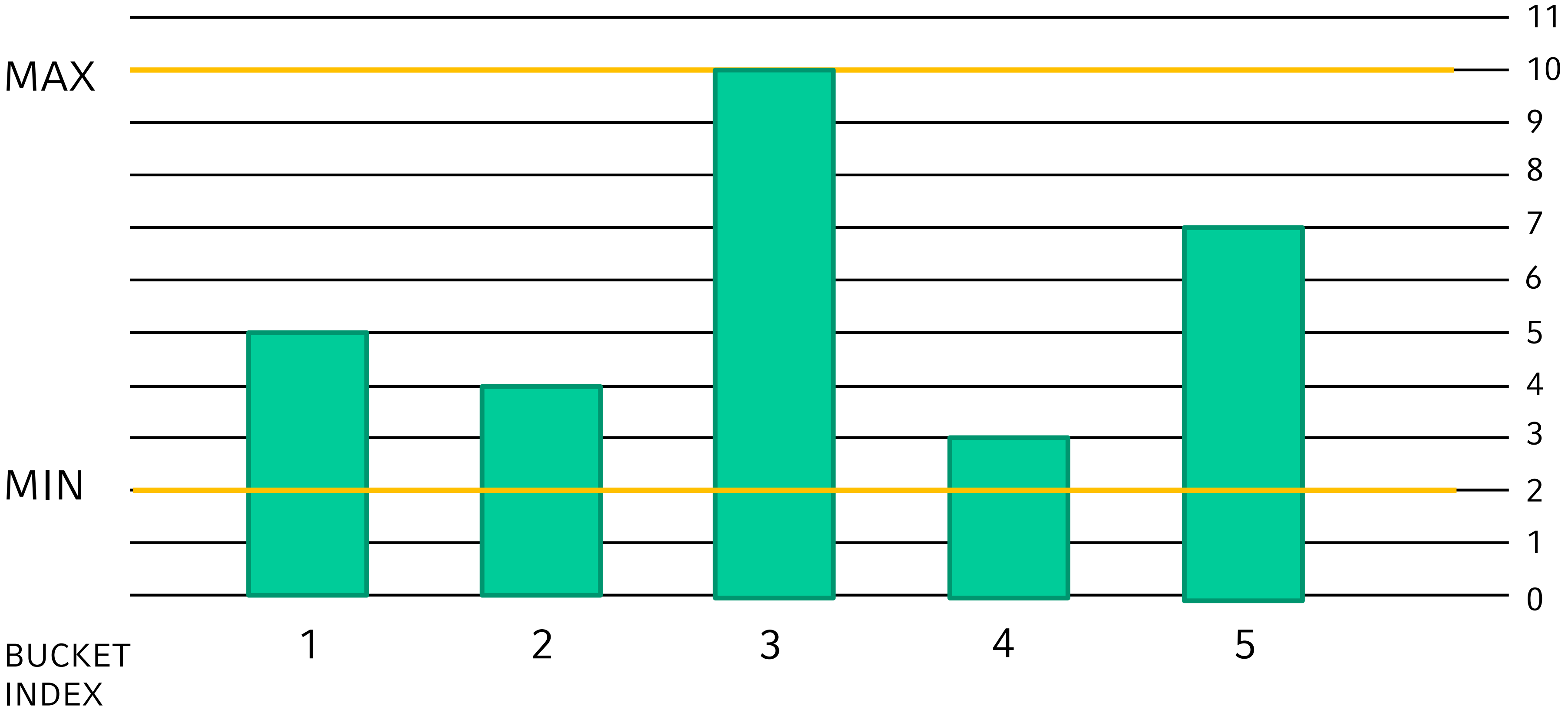


Assignment 6-3

Sequence = 3, 1, 3, 5, 2, 3, 4, 1, 5, 3

Mode: INSERTING

INSERT 2



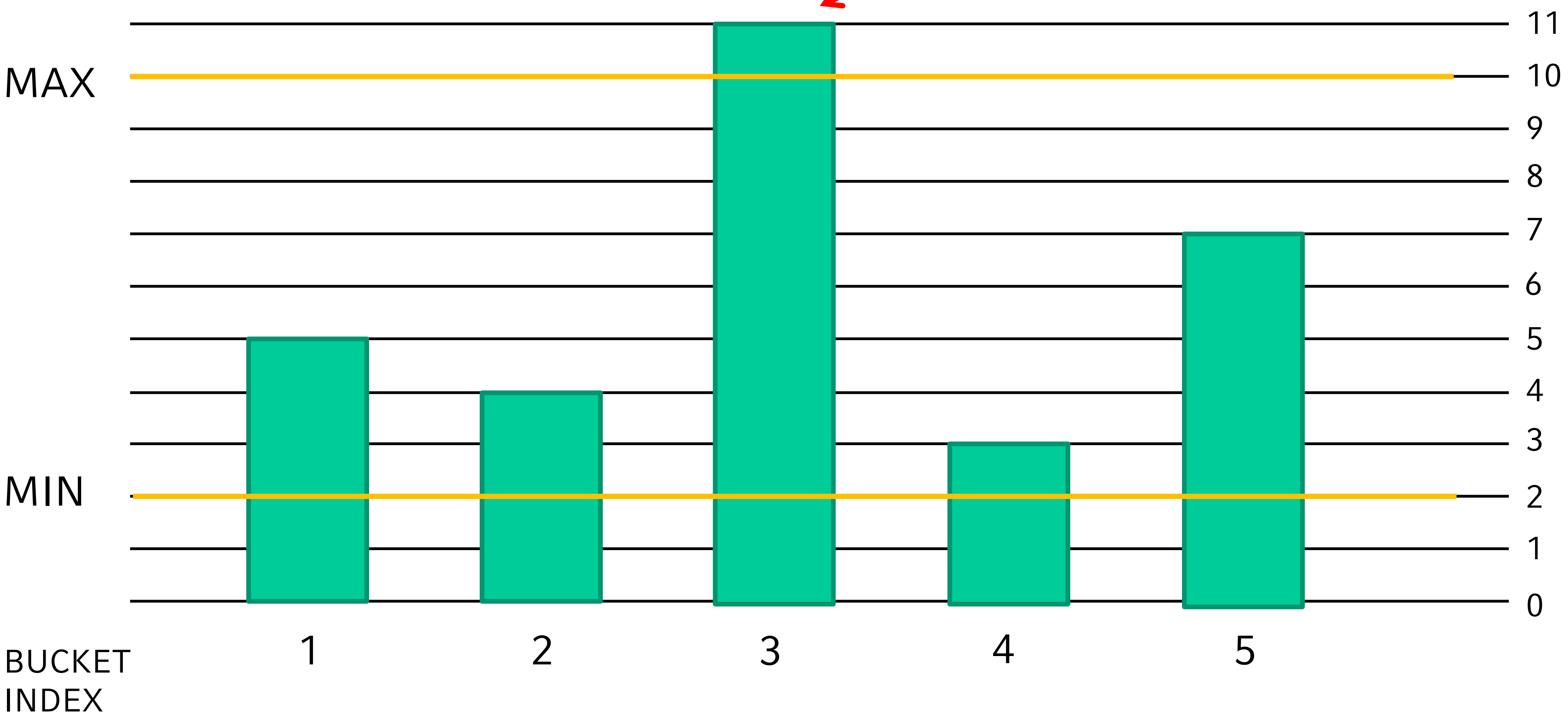
Assignment 6-3

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

Mode: INSERTING

INSERT 3

Threshold exceeded! → STOP



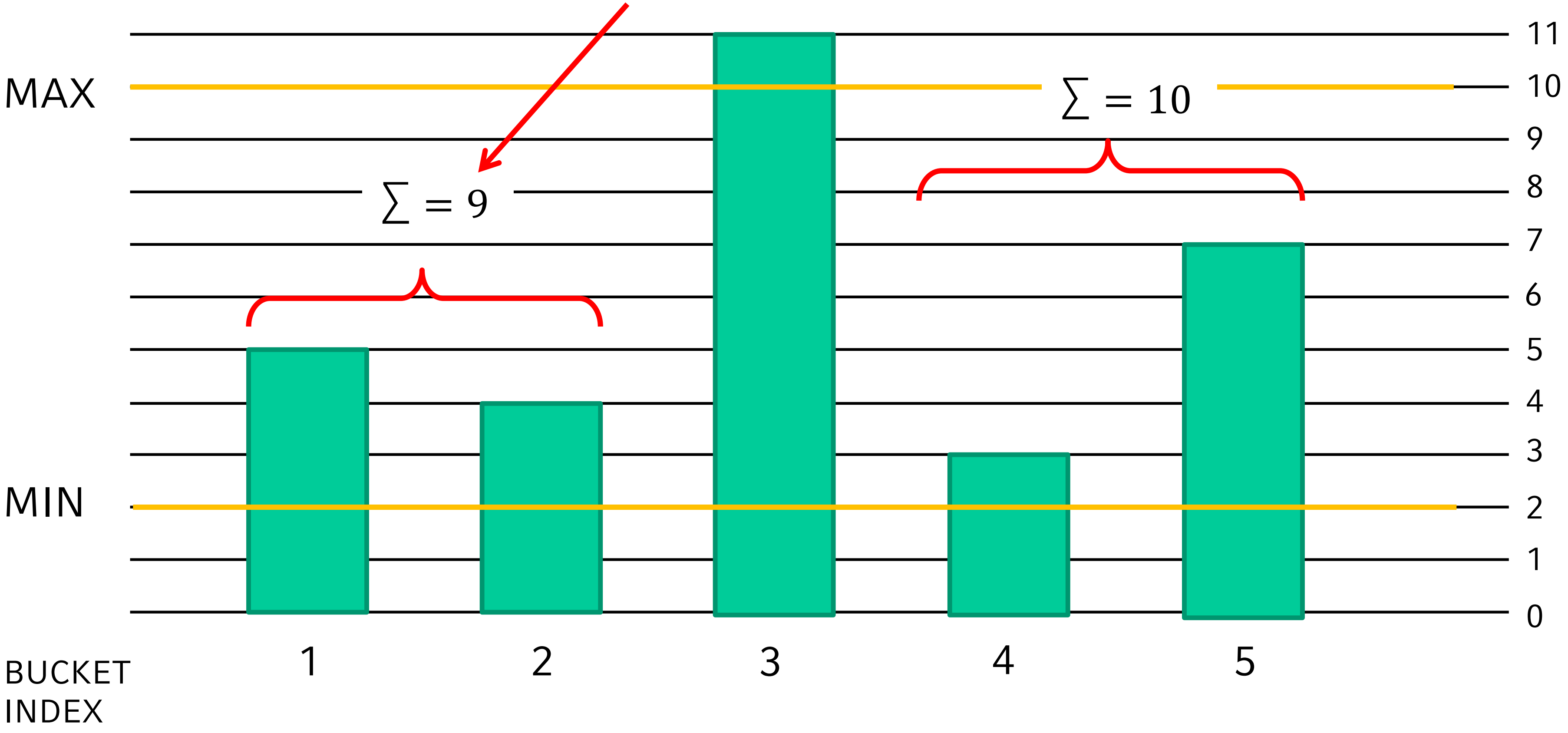
Assignment 6-3

Split & Merge

Sequence = 3, 1, 3, 5, 2, 3, 4, 1, 5, 3

Mode: INSERTING

Take the two consecutive buckets with the lowest overall sum of sizes



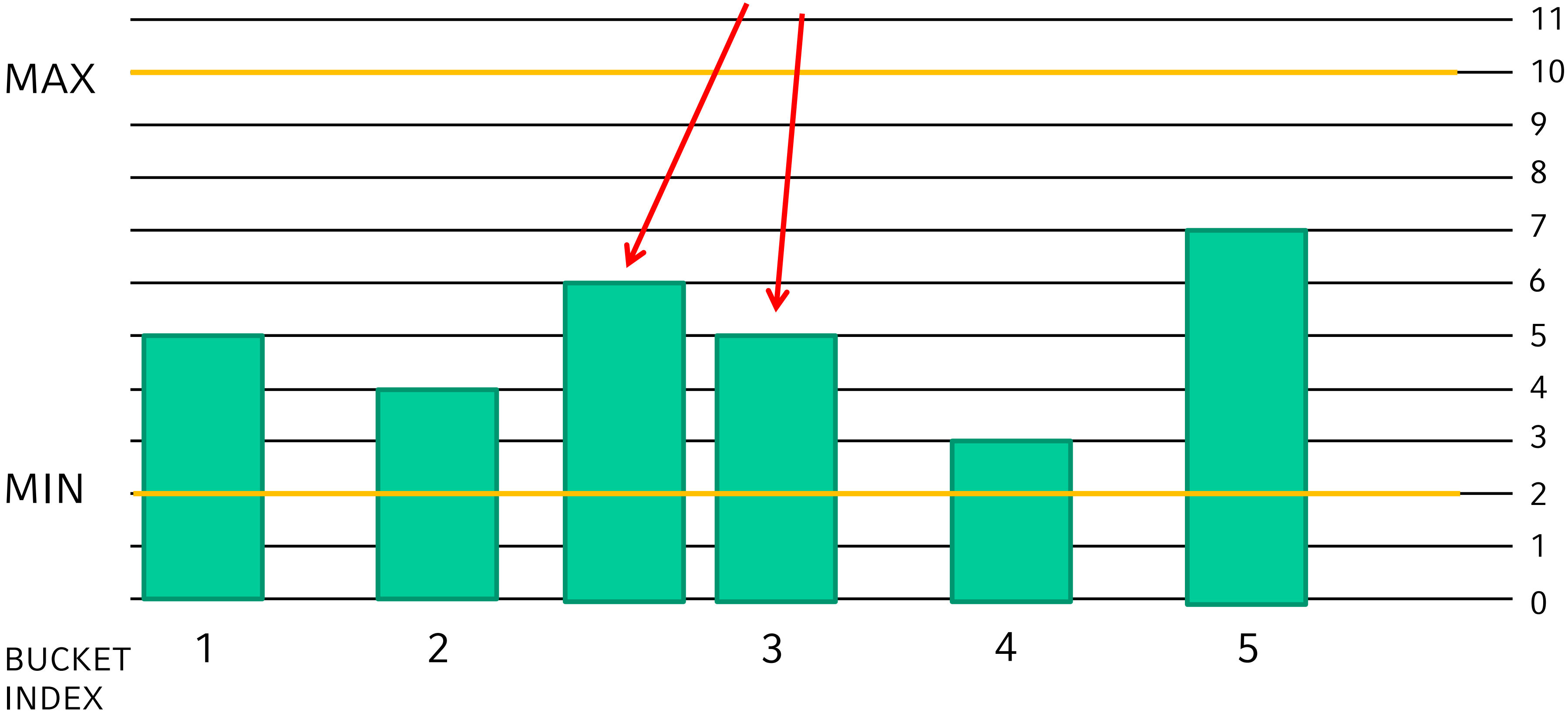
Assignment 6-3

Split & Merge

Sequence = 3, 1, 3, 5, 2, 3, 4, 1, 5, 3

Mode: INSERTING

Split bucket 3 [size 11] (in half, floor function for bucket 3 if bucket size odd)



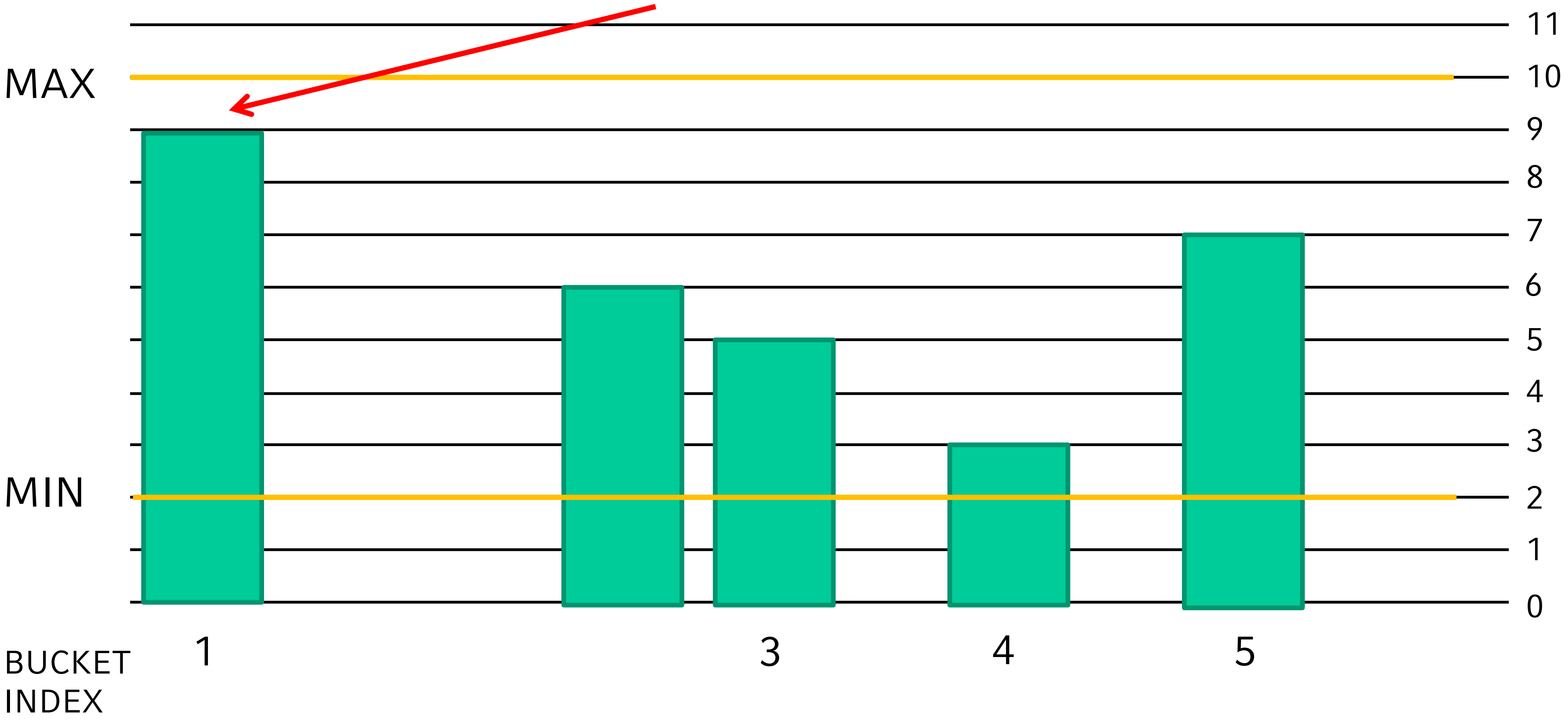
Assignment 6-3

Split & Merge

Sequence = 3, 1, 3, 5, 2, 3, 4, 1, 5, 3

Mode: INSERTING

Merge buckets 1 [size 5] and 2 [size 4] to a new bucket 1



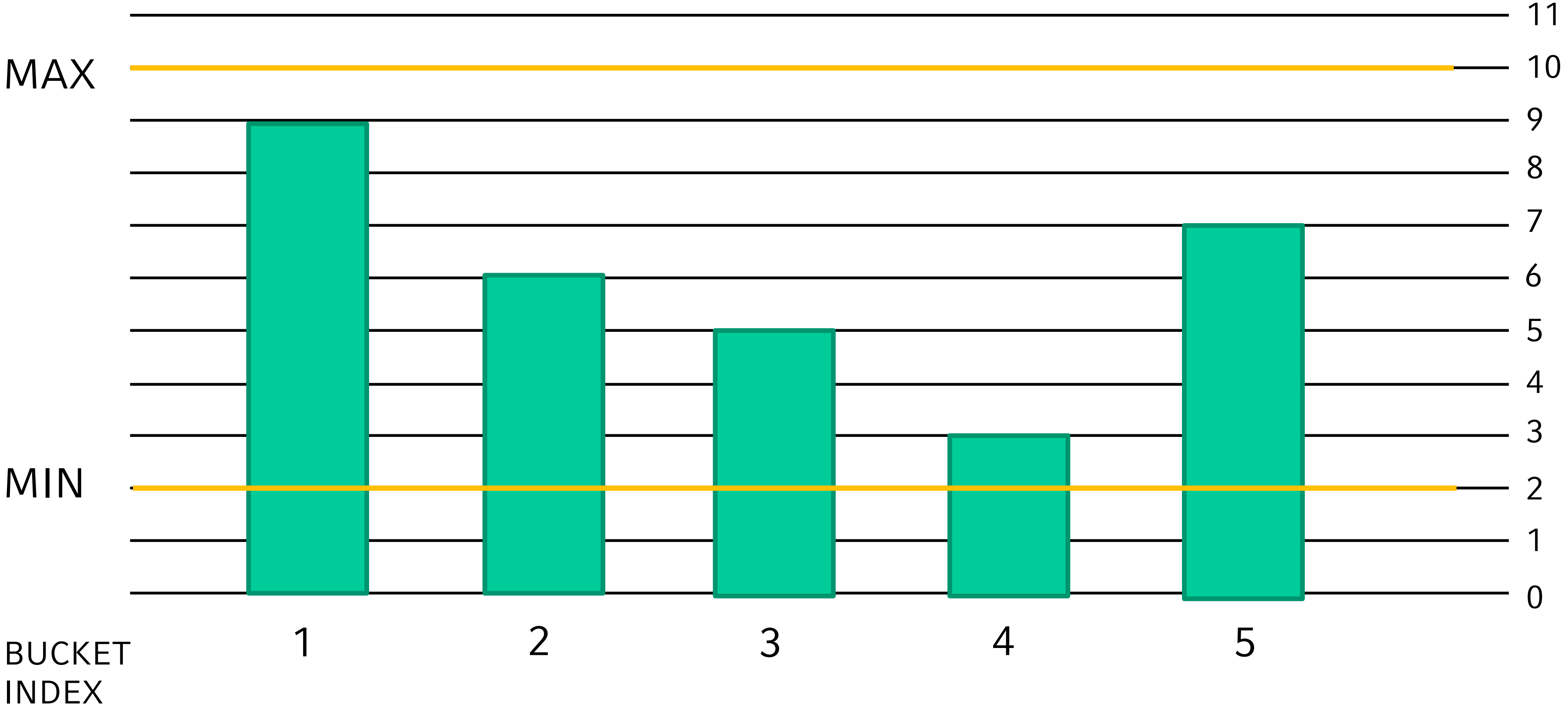
Assignment 6-3

Split & Merge

Sequence = 3, 1, 3, 5, 2, 3, 4, 1, 5, 3

Mode: INSERTING

Assign new indices!

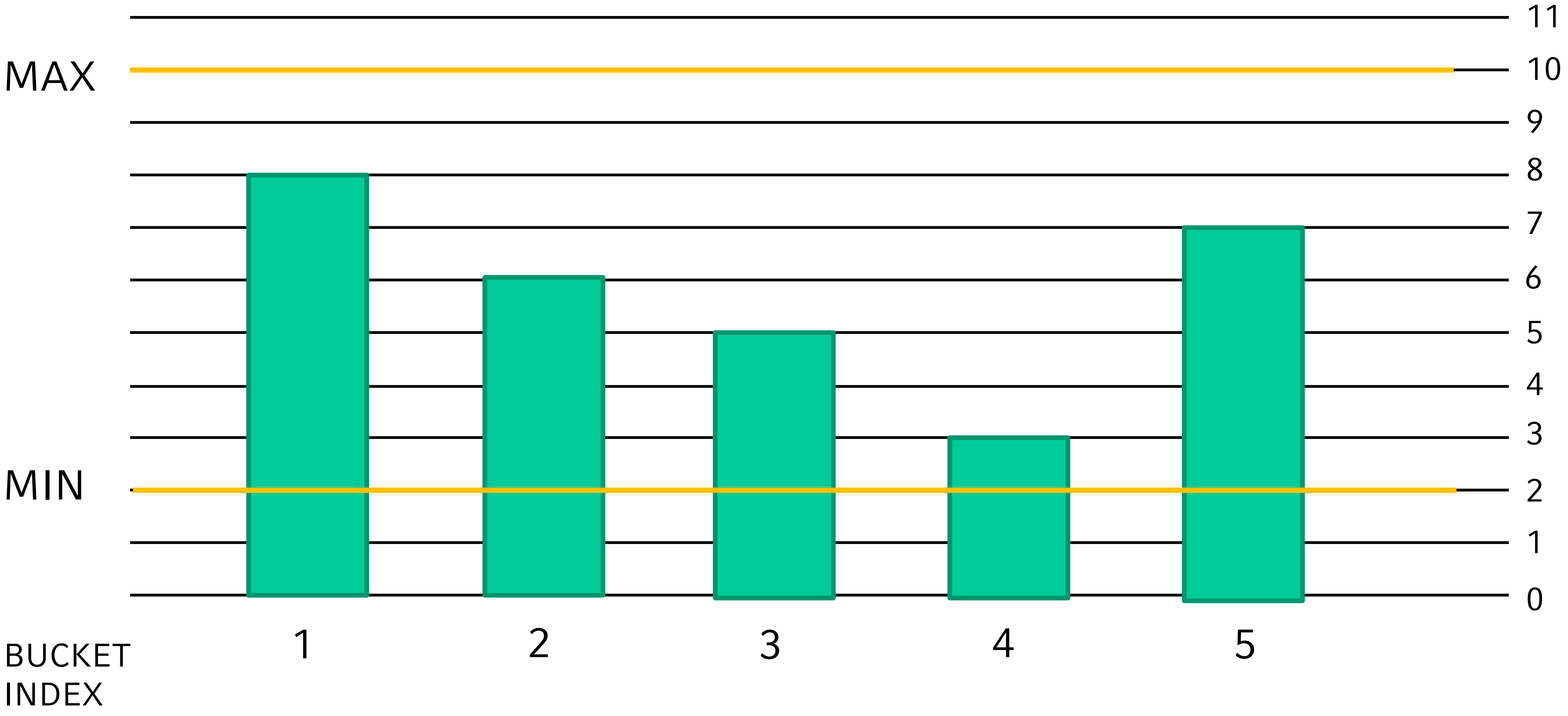


Assignment 6-3

Sequence = 1, 3, 4, 5, 4, 3, 2, 5, 1, 2

Mode: DELETING

DELETE 1

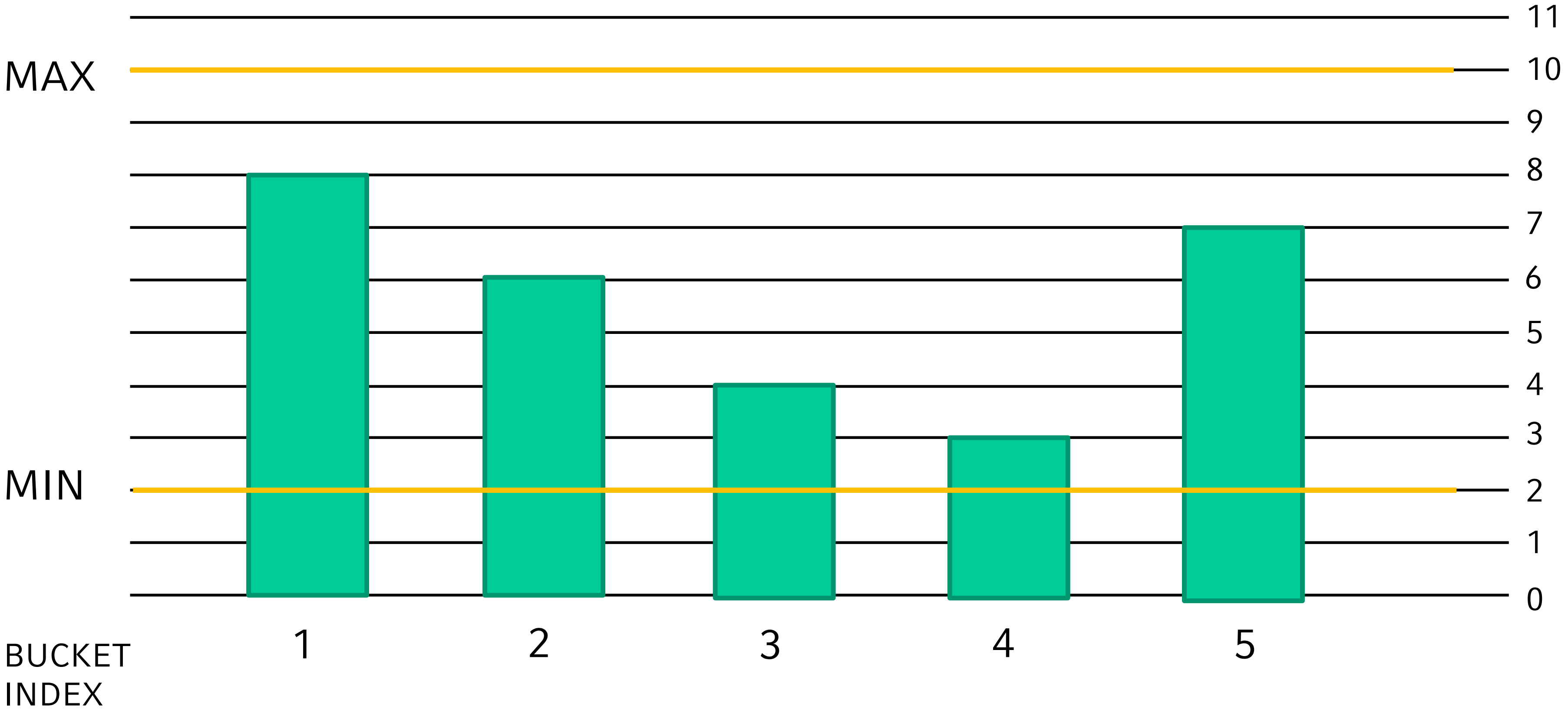


Assignment 6-3

Sequence = 1, 3, 4, 5, 4, 3, 2, 5, 1, 2

Mode: DELETING

DELETE 3

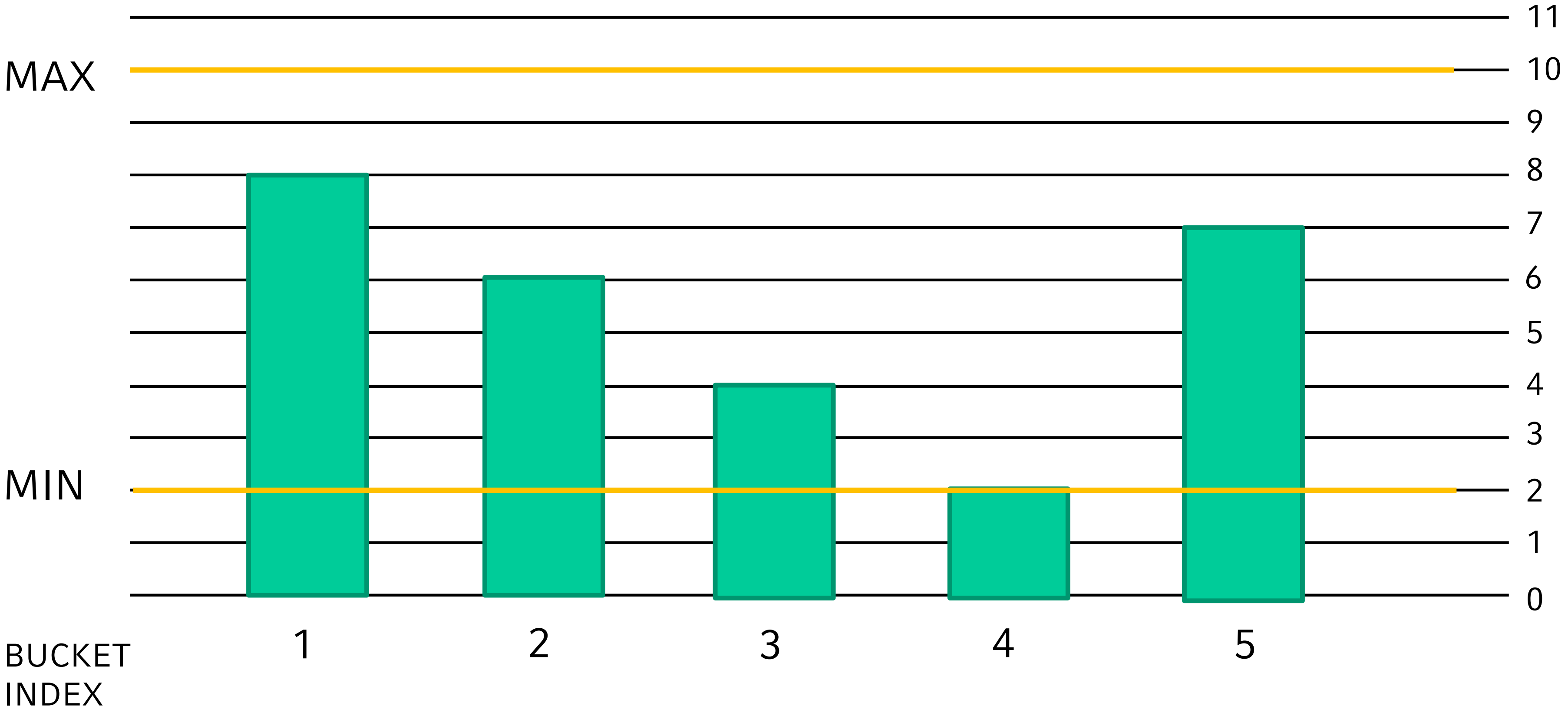


Assignment 6-3

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

Mode: DELETING

DELETE 4

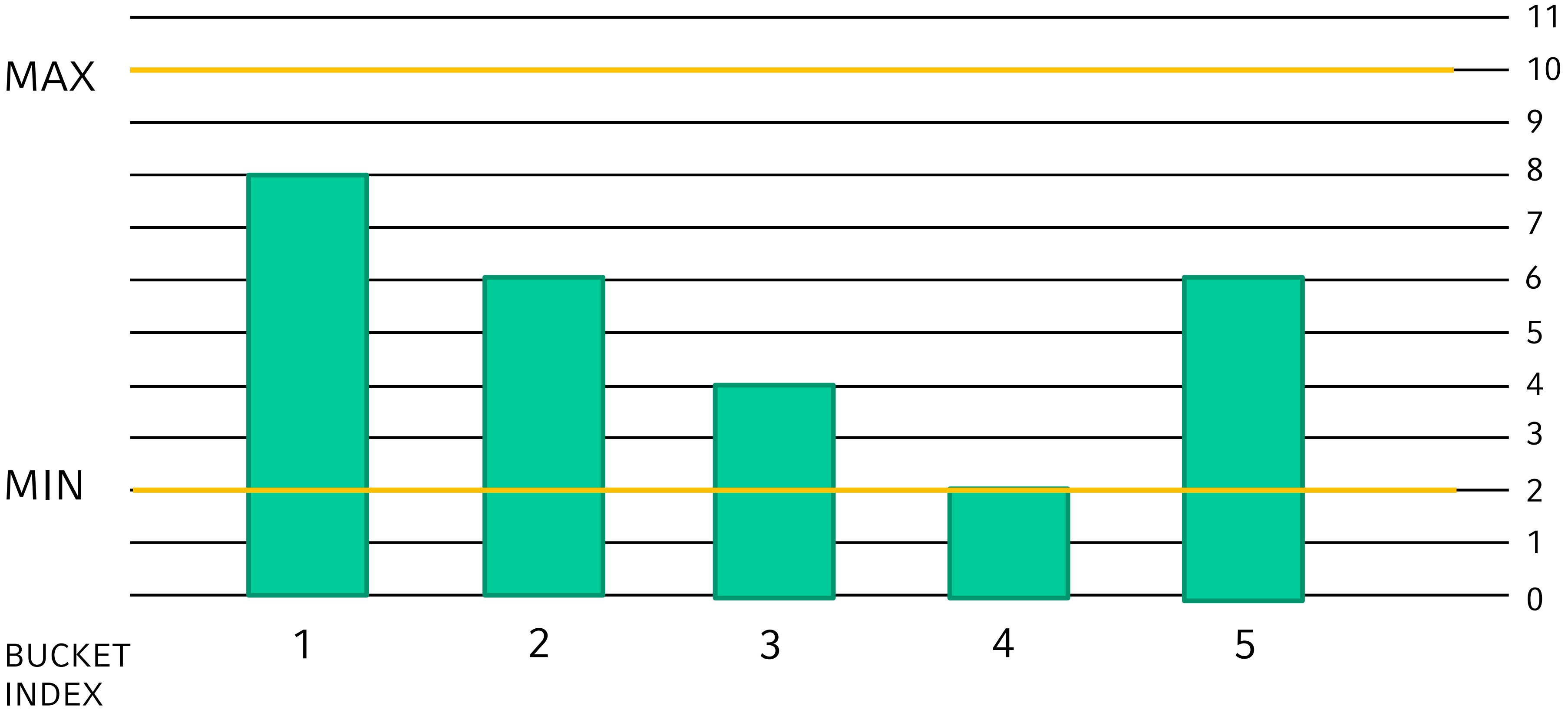


Assignment 6-3

Sequence = 1, 3, 4, 5, 4, 3, 2, 5, 1, 2

Mode: DELETING

DELETE 5



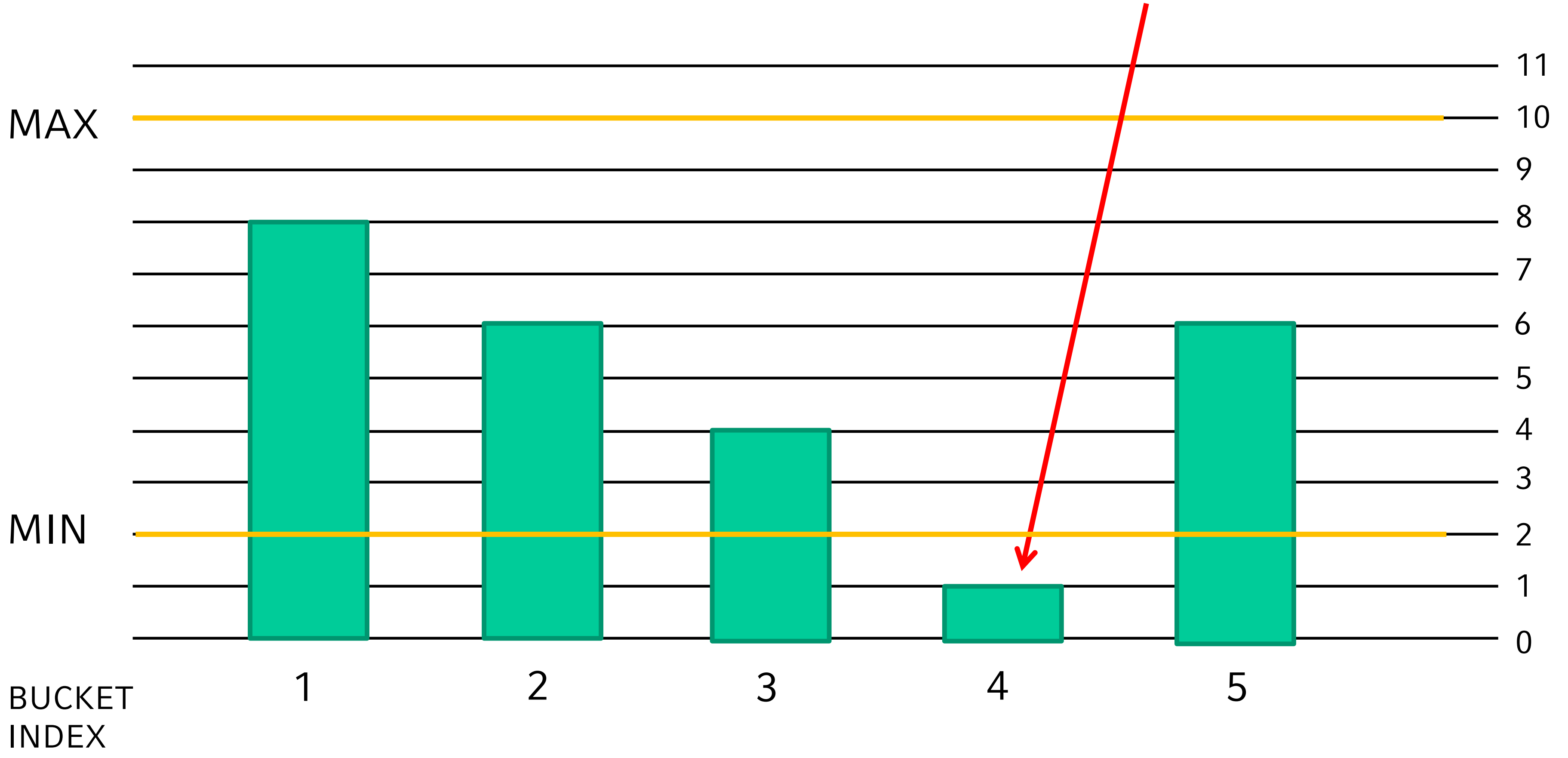
Assignment 6-3

Sequence = 1, 3, 4, 5, 4, 3, 2, 5, 1, 2

Mode: DELETING

DELETE 4

Threshold underflow! → STOP



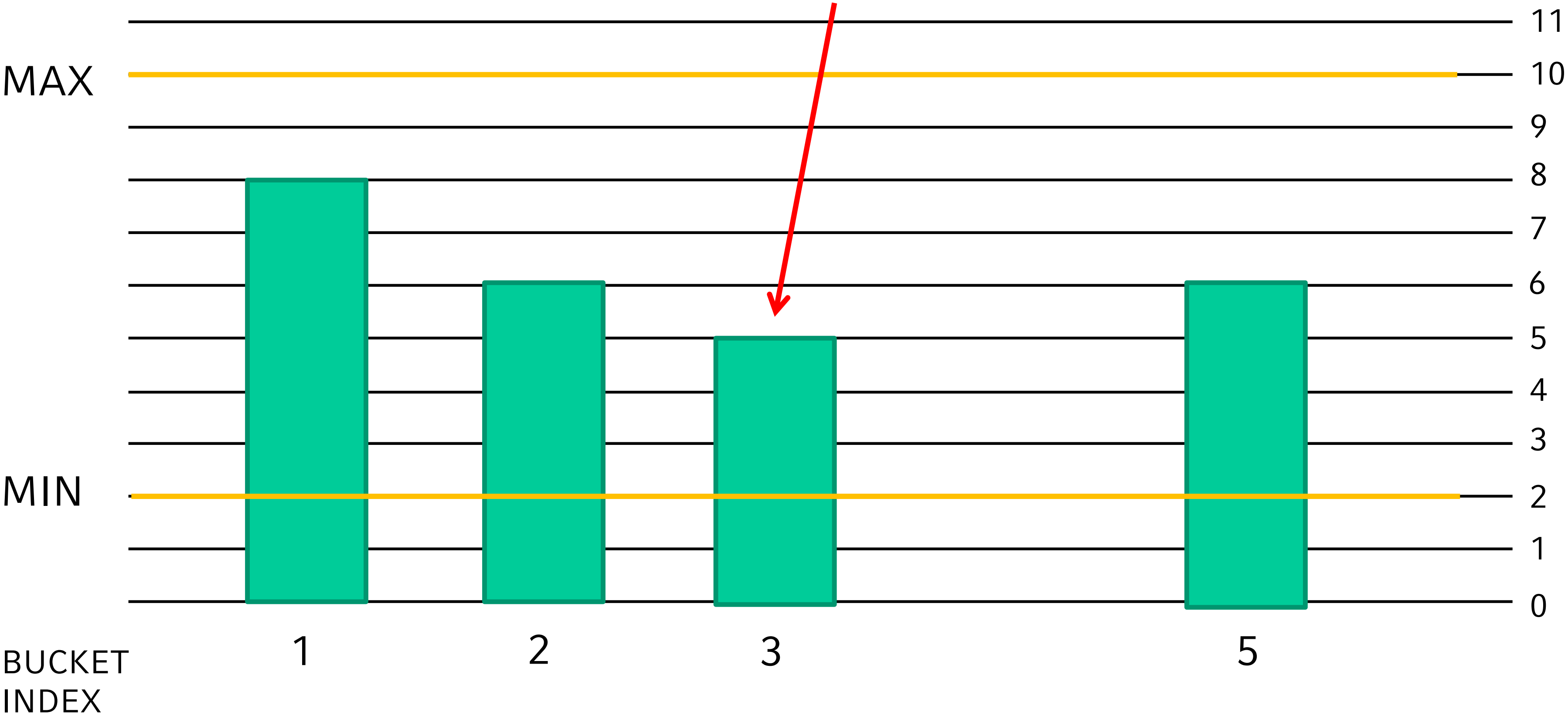
Assignment 6-3

Sequence = 1, 3, 4, 5, 4, 3, 2, 5, 1, 2

Merge & Split

Mode: DELETING

Merge bucket 4 [size 1] with the neighbor bucket that has the smallest size (bucket 3 [size 4])



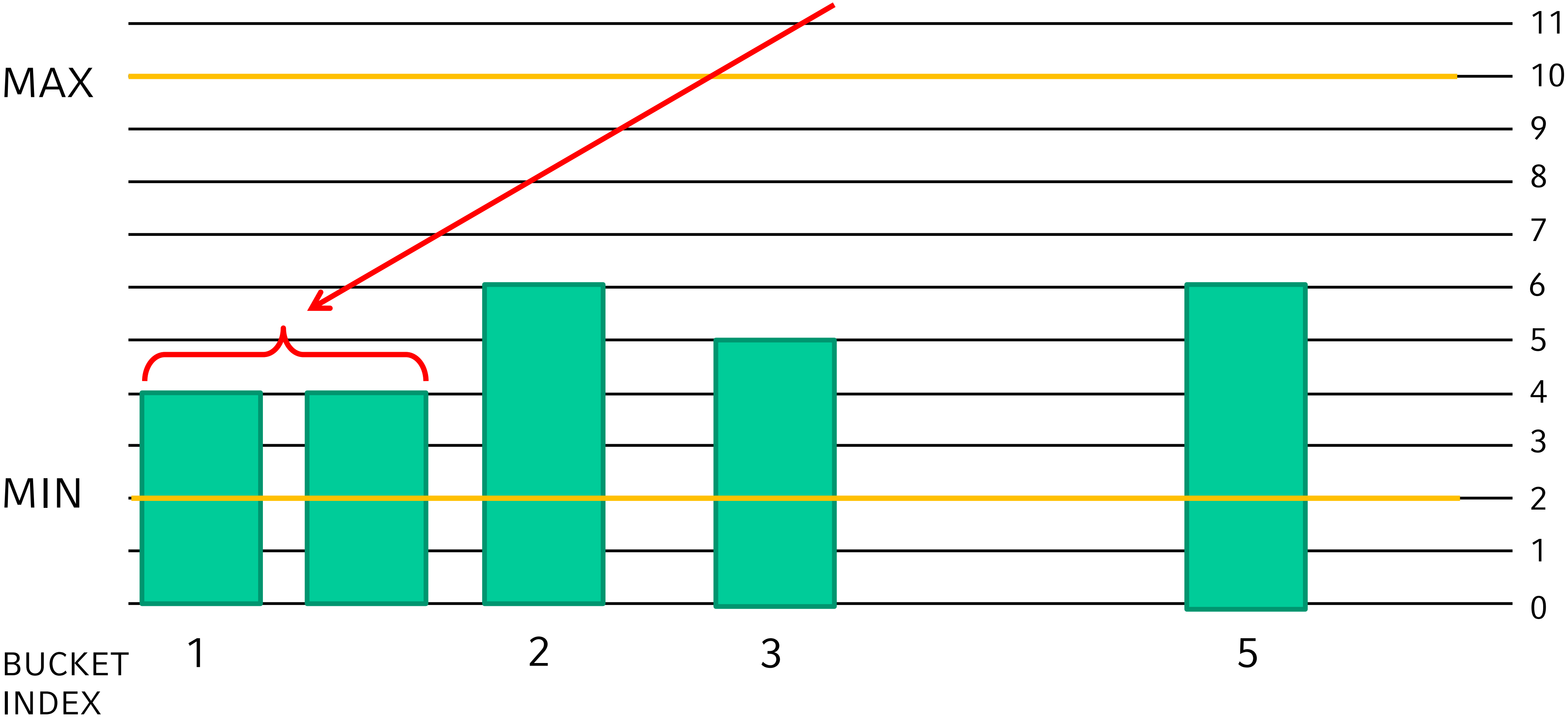
Assignment 6-3

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

Merge & Split

Mode: DELETING

Split bucket with the largest size (bucket 1) in half (8 → 4 , 4)



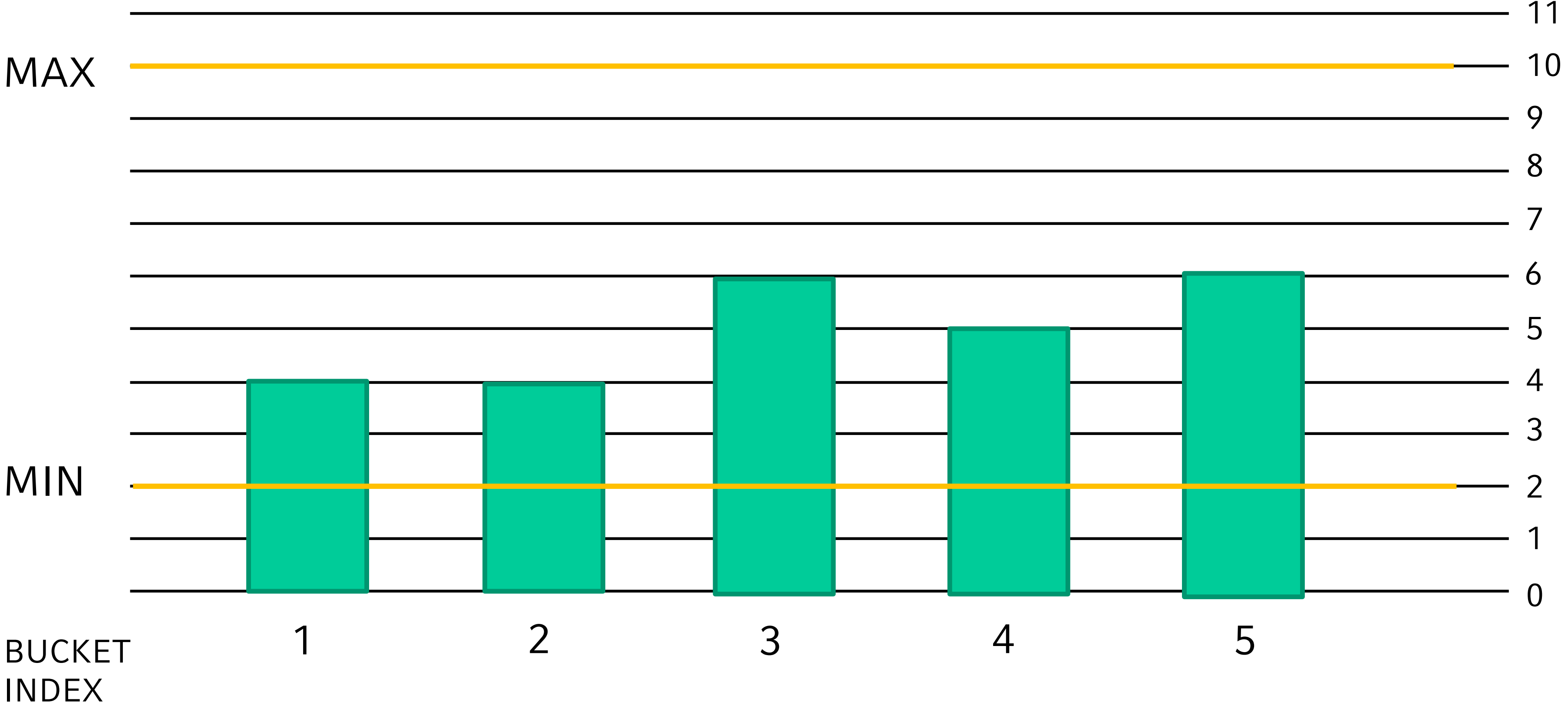
Assignment 6-3

Sequence = 1, 3, 4, 5, 4, 3, 2, 5, 1, 2

Mode: DELETING

Split & Merge

Assign new indices



Assignment 6-4

CUSUM - CUmulative SUM

Purpose: Change detection on data streams

Core idea: Observe cumulative sum of instances of a random variable

Detection mechanism: If the normalized mean of the input data differs from 0 by an threshold α

Assignment 6-4

The formula for detecting changes is:

$$\boxed{G_t \coloneqq max(0, G_{t-1} - \omega_t + x_t)}$$

where:

G_t: cumulative sum

ω_t: assigned weights

x_t: next sample from a data stream S

The original CUSUM algorithm detects positive changes. In order to detect also negative changes we modify the equation above to:

$$\boxed{G_t \coloneqq (G_{t-1} - \omega_t + x_t)}$$

Assignment 6-4

Given:

Sequence S = (2,3,7,4,0,2,5,6,8,7)

Mean $\omega = 3$

Threshold $\alpha = 8$

t	$x_t - \omega$	G_t
0	-	0
1	-1	-1
2	0	-1
3	4	3
4	1	4
5	-3	1
6	-1	0
7	2	2
8	3	5
9	5	10
10	4	4

$G_t > \alpha$

$10 > 8$

Change detected between t=8 and t=9

```
if $G_t > \alpha$ then
    report change at time $t$
    $G_t := 0$
```