DATABASE
SYSTEMS
GROUP

LMU

Big Data Management and Analytics
Assignment 9

Assignment 9-1

(a) k-Bucket histograms:

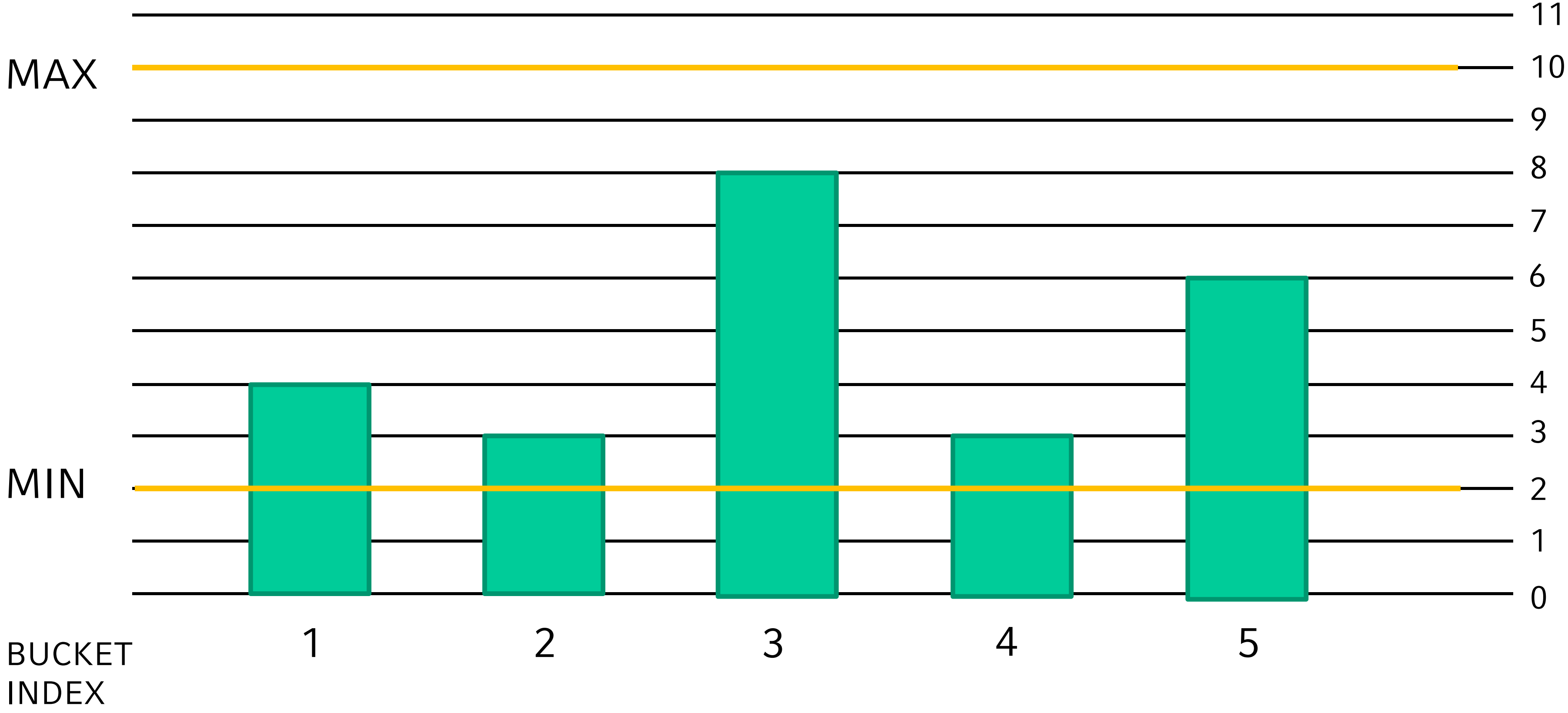
- Histogram consists constantly of k=5 buckets
- Upper threshold per bucket MAX = 10
- Lower threshold per bucket MIN = 2

DATABASE SYSTEMS GROUP

Assignment 9-1

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

Mode: INSERTING

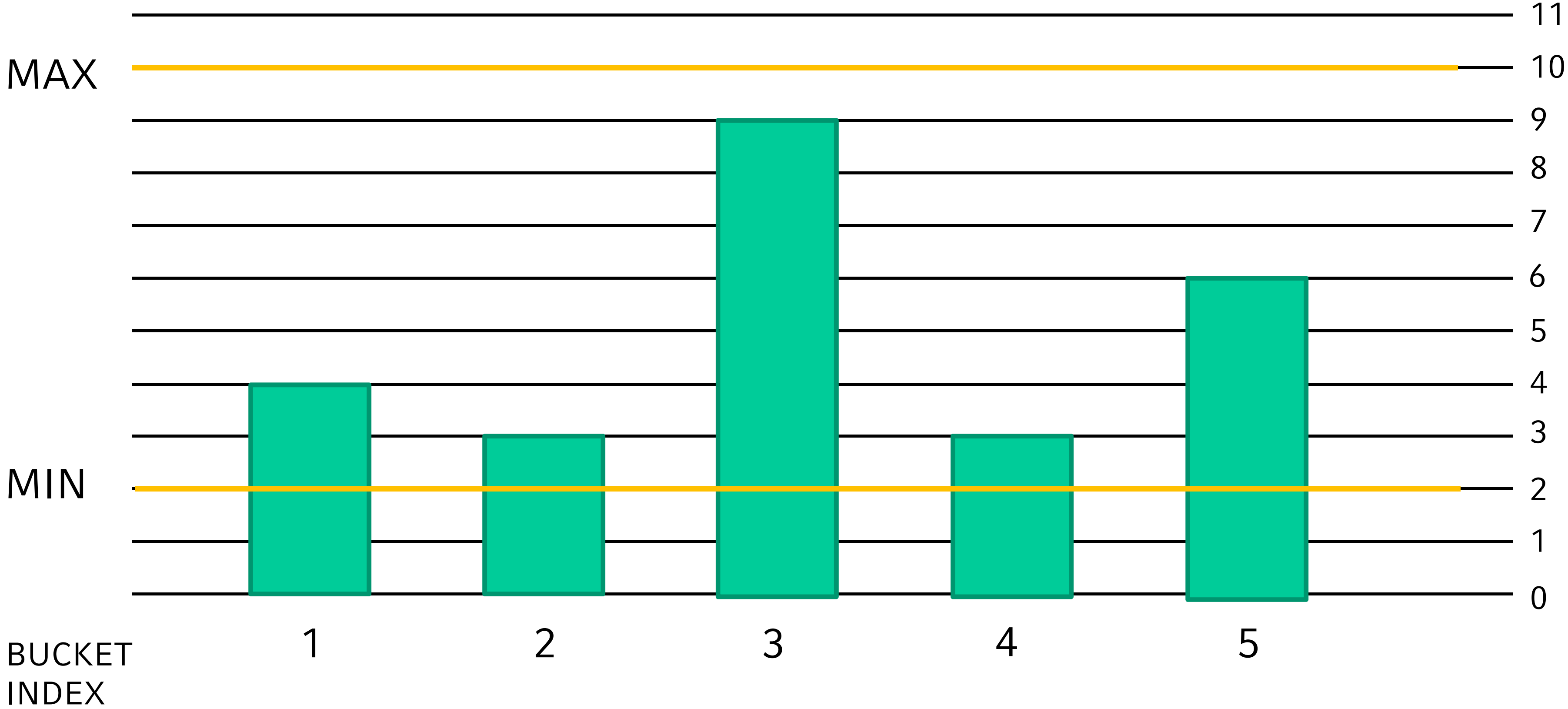


Assignment 9-1

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

INSERT 3

Mode: INSERTING

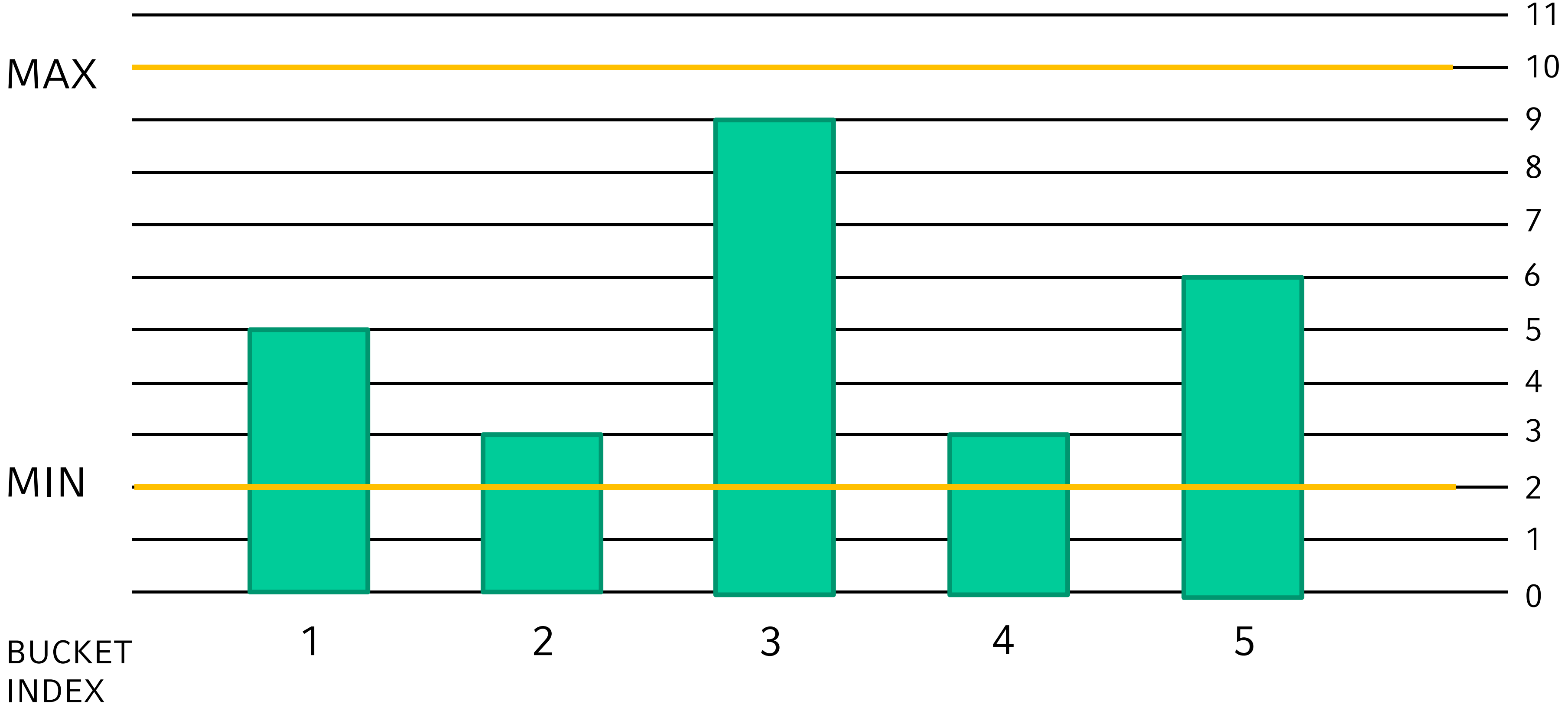


Assignment 9-1

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

Mode: INSERTING

INSERT 1

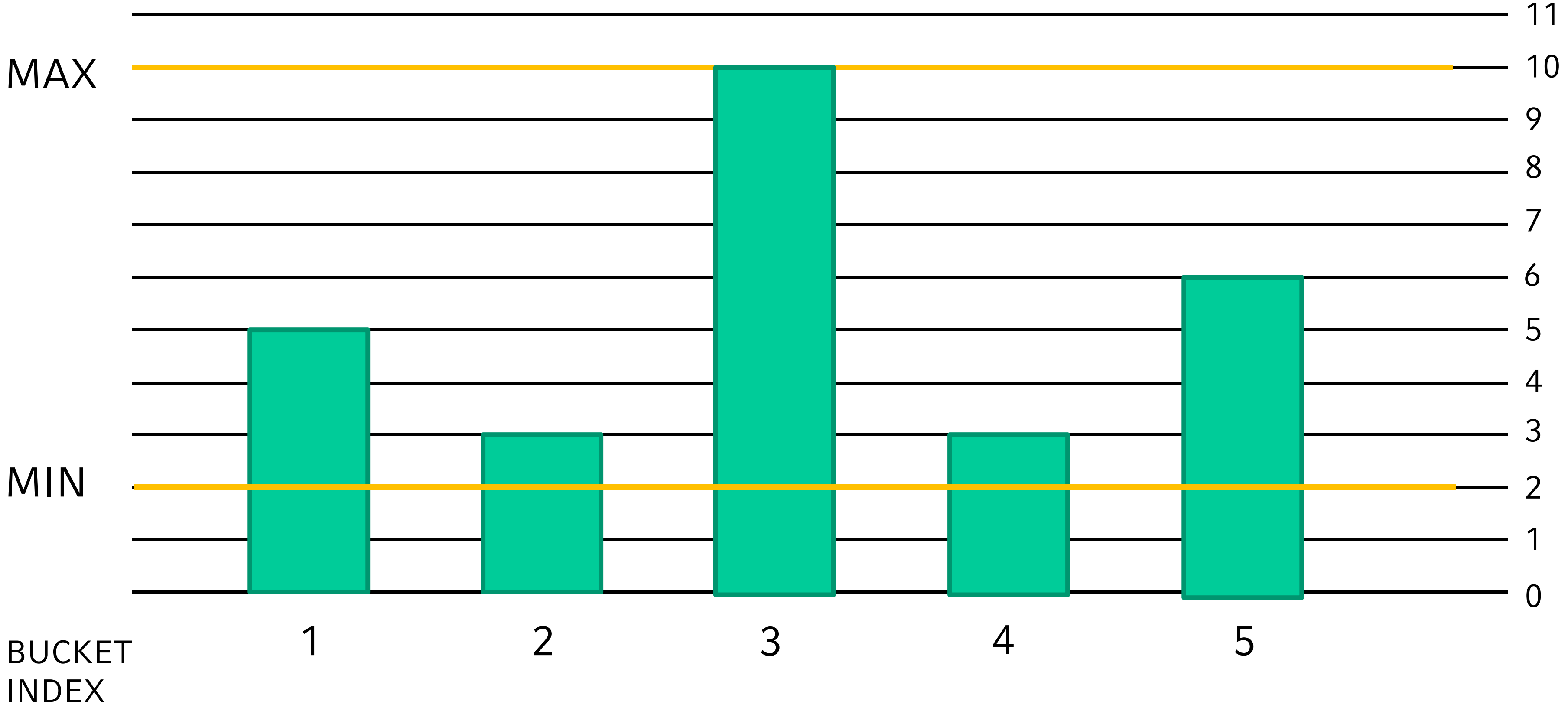


Assignment 9-1

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

Mode: INSERTING

INSERT 3

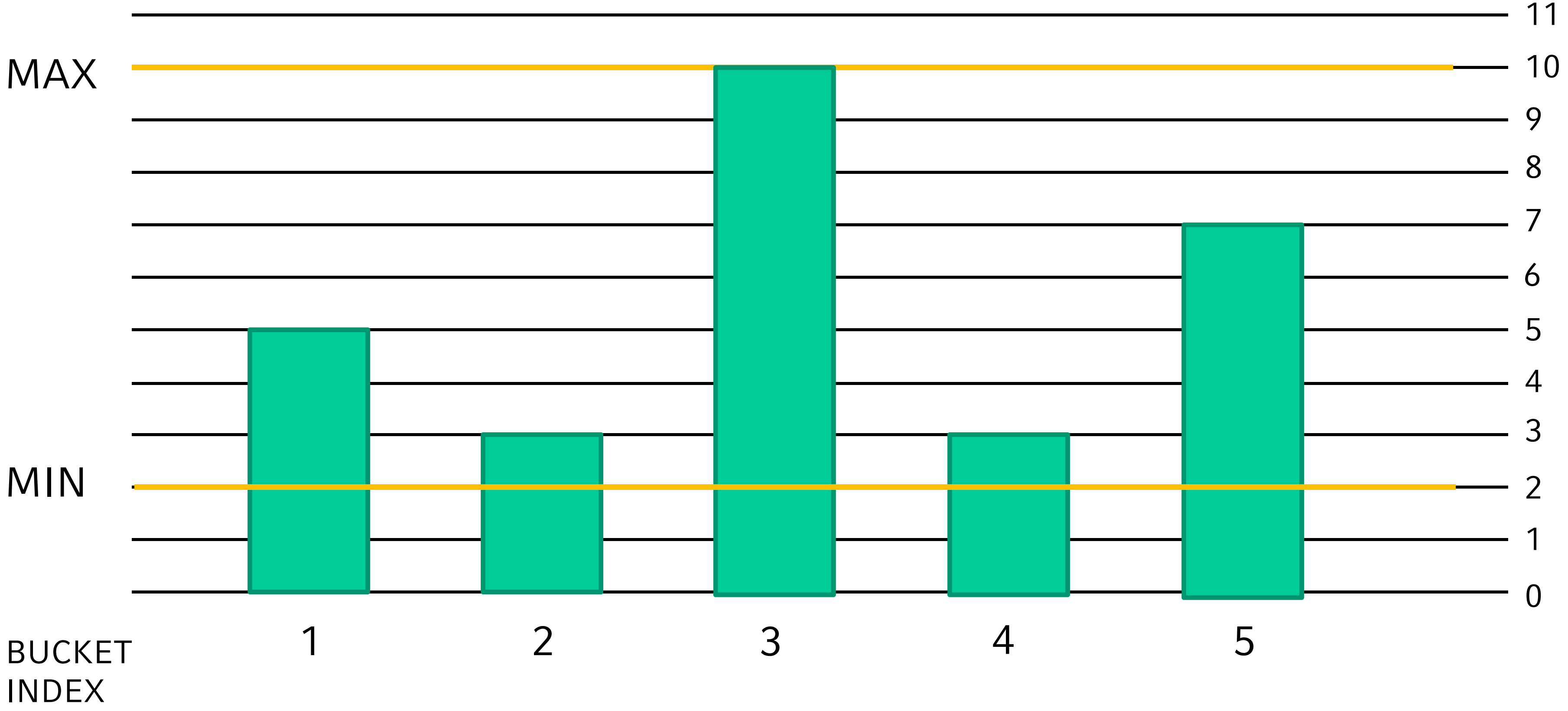


Assignment 9-1

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

Mode: INSERTING

INSERT 5

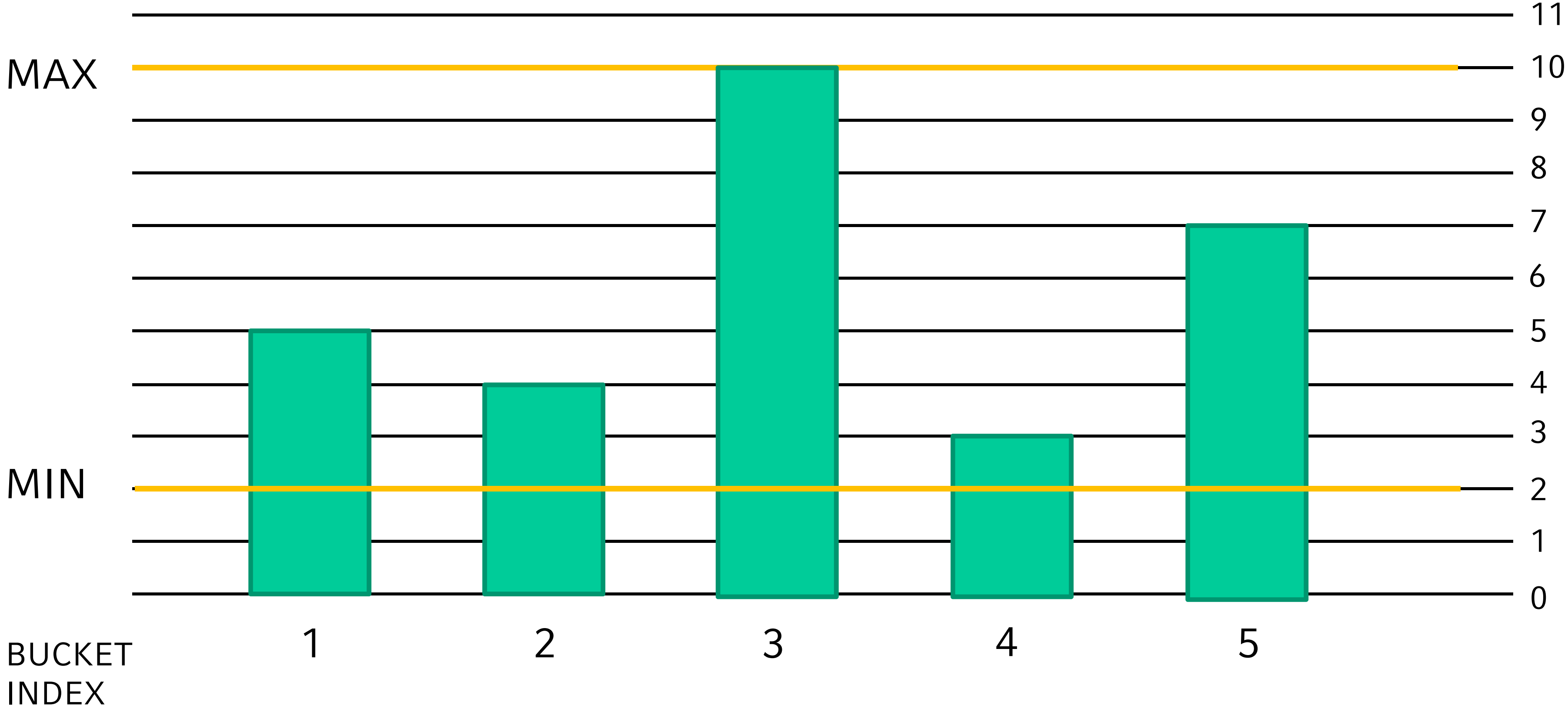


Assignment 9-1

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

Mode: INSERTING

INSERT 2



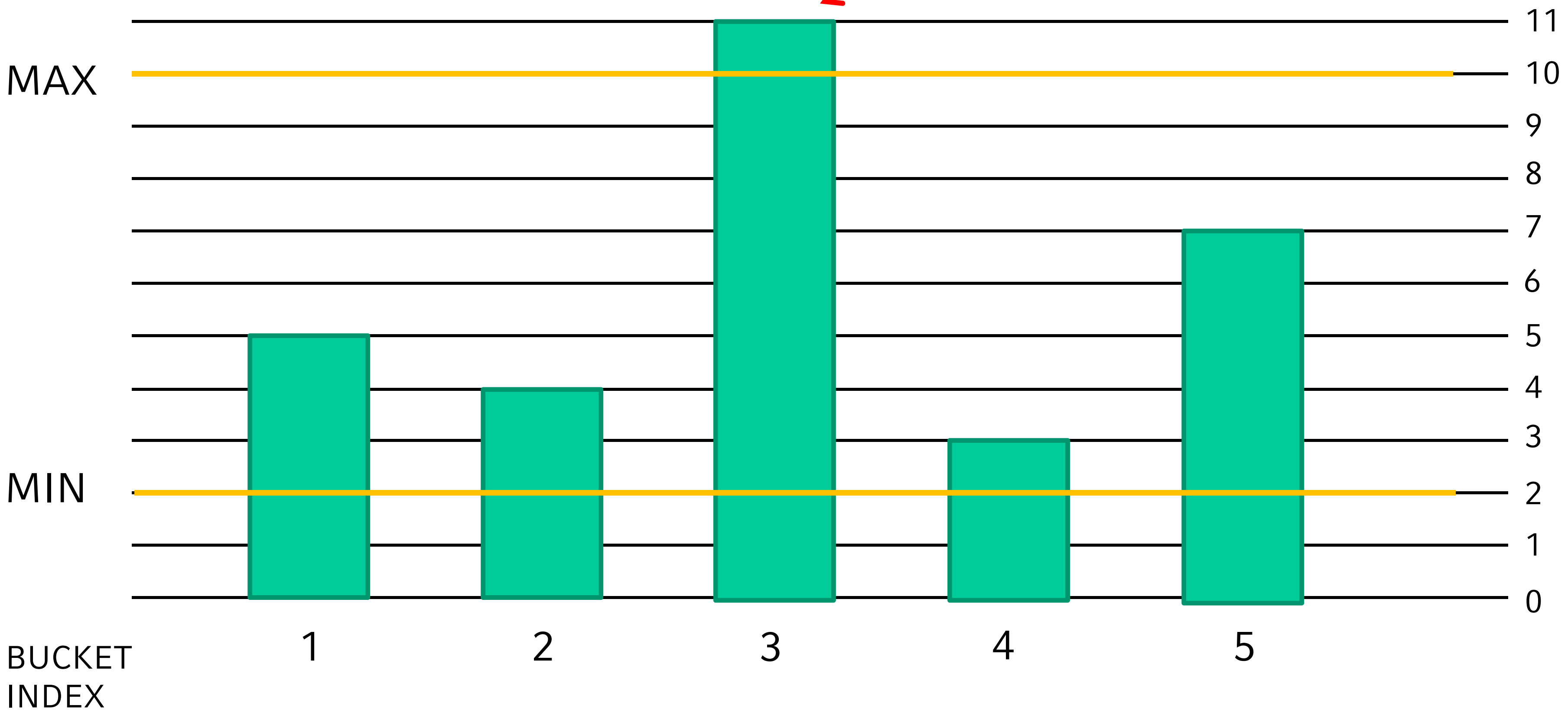
Assignment 9-1

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

Mode: INSERTING

INSERT 3

Threshold exceeded! → STOP

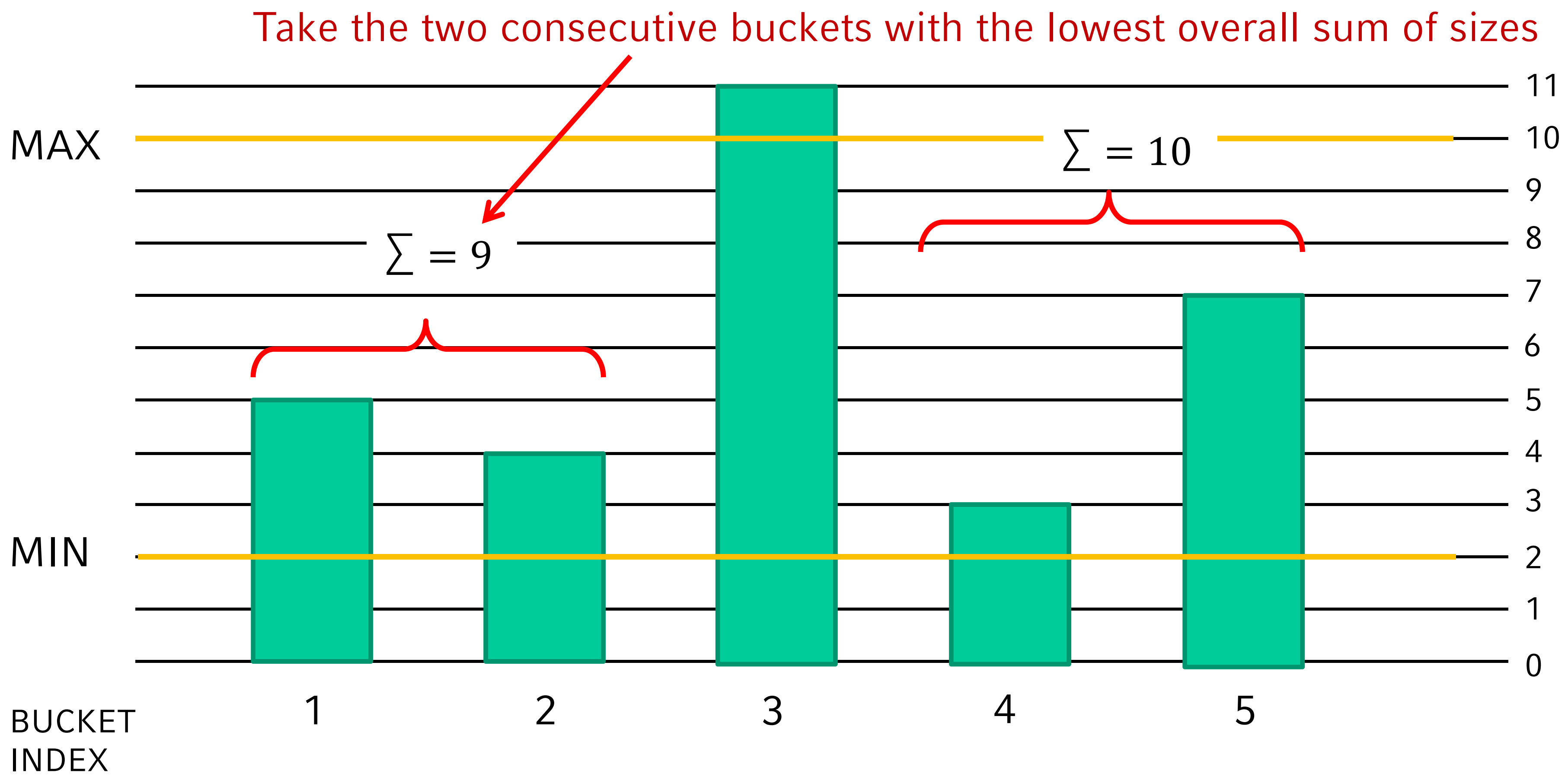


Assignment 9-1

Split & Merge

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

Mode: INSERTING



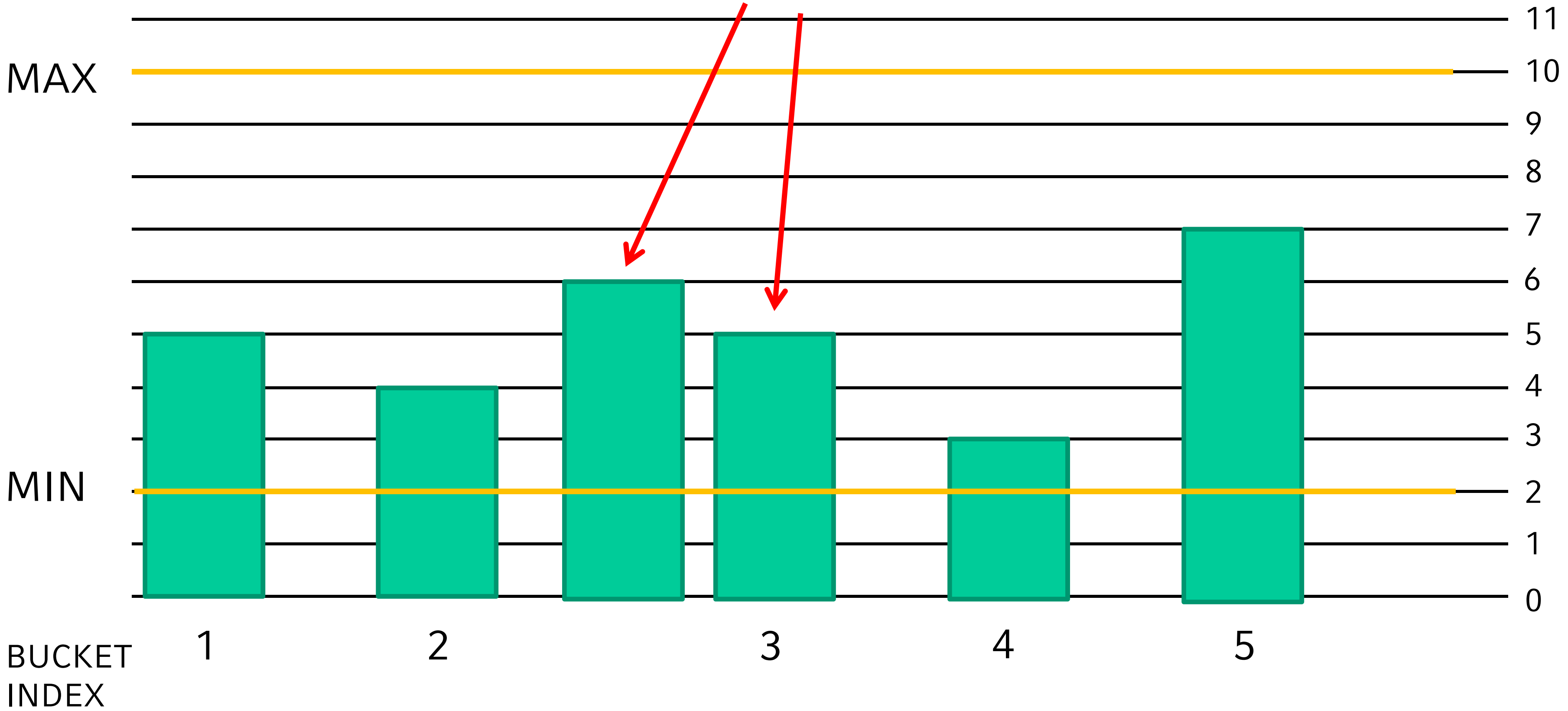
Assignment 9-1

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)



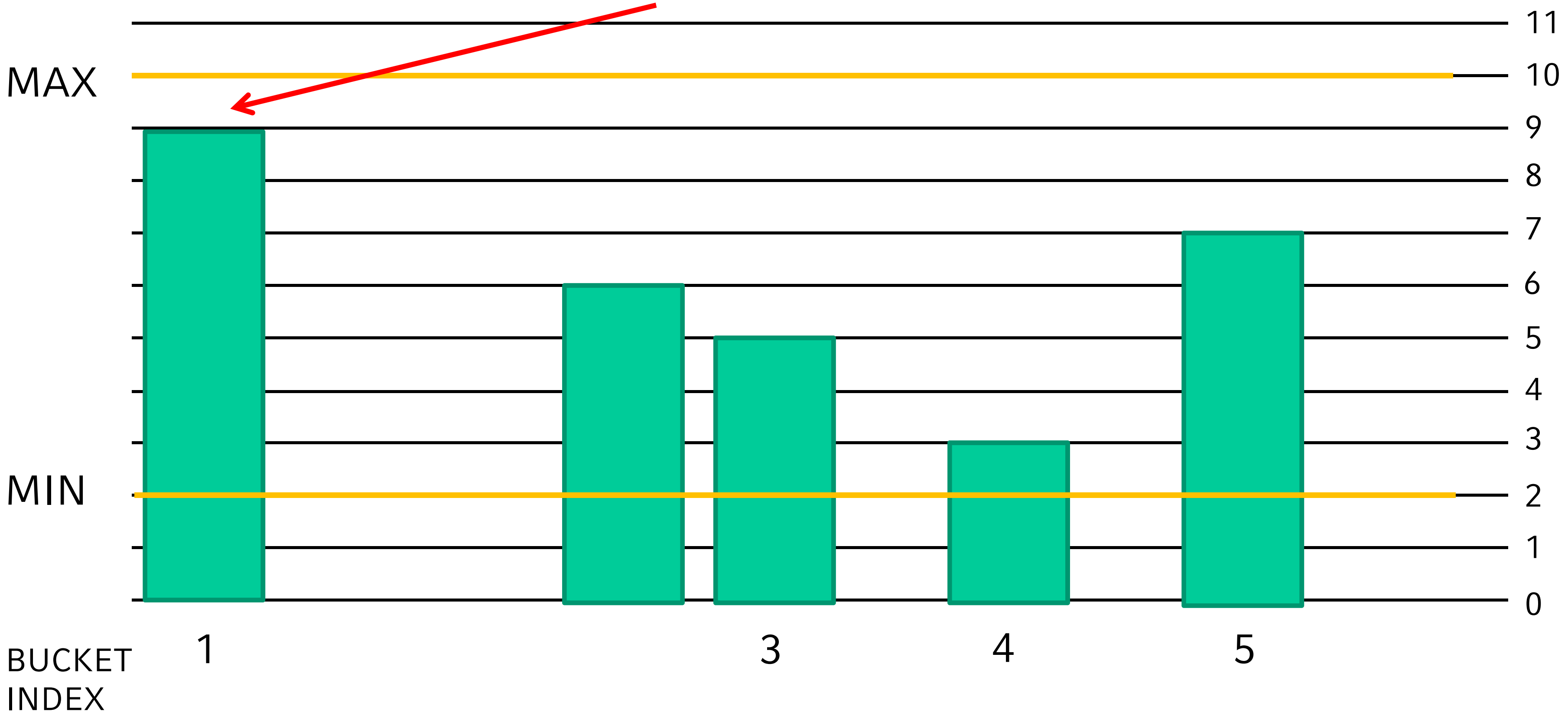
Assignment 9-1

Split & Merge

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

Mode: INSERTING

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



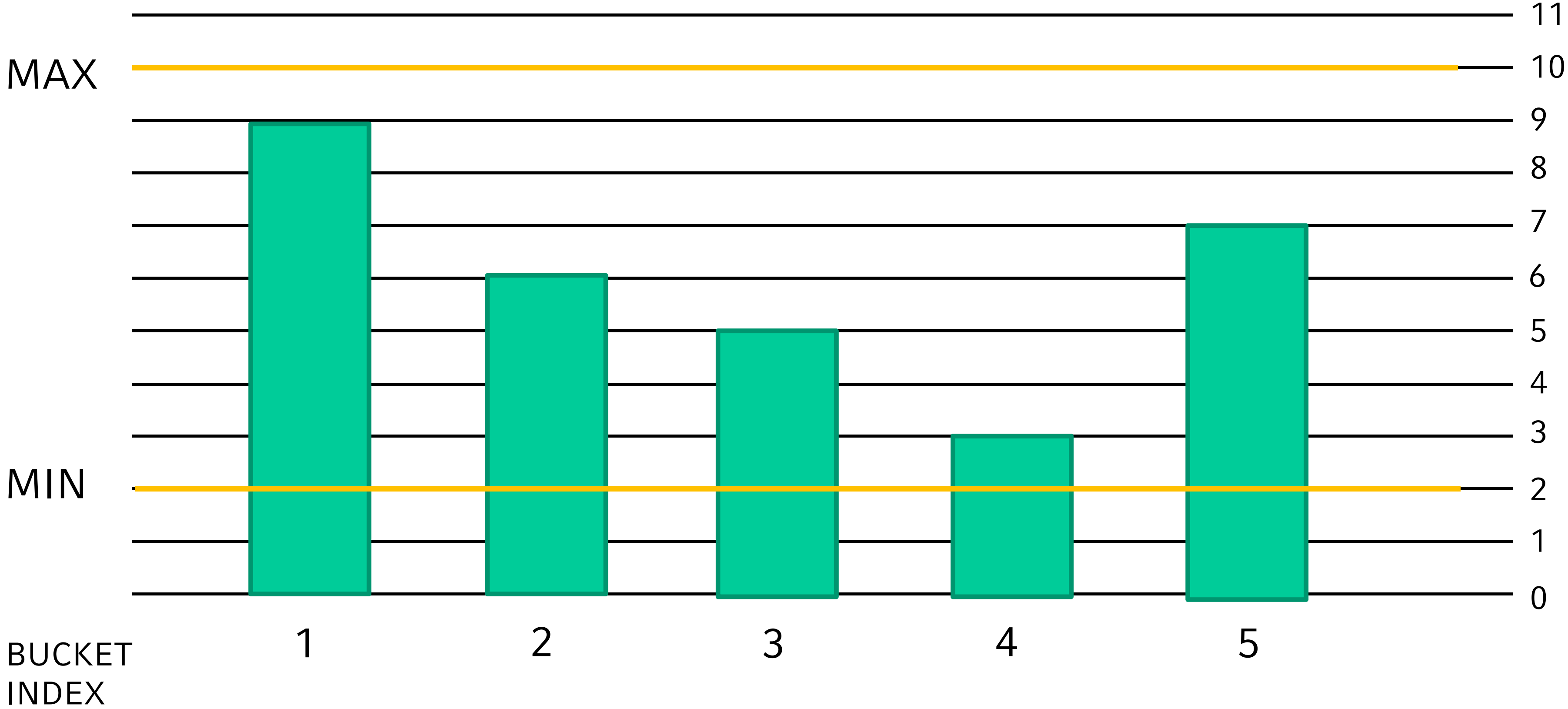
Assignment 9-1

Split & Merge

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

Mode: INSERTING

Assign new indices!



DATABASE SYSTEMS GROUP

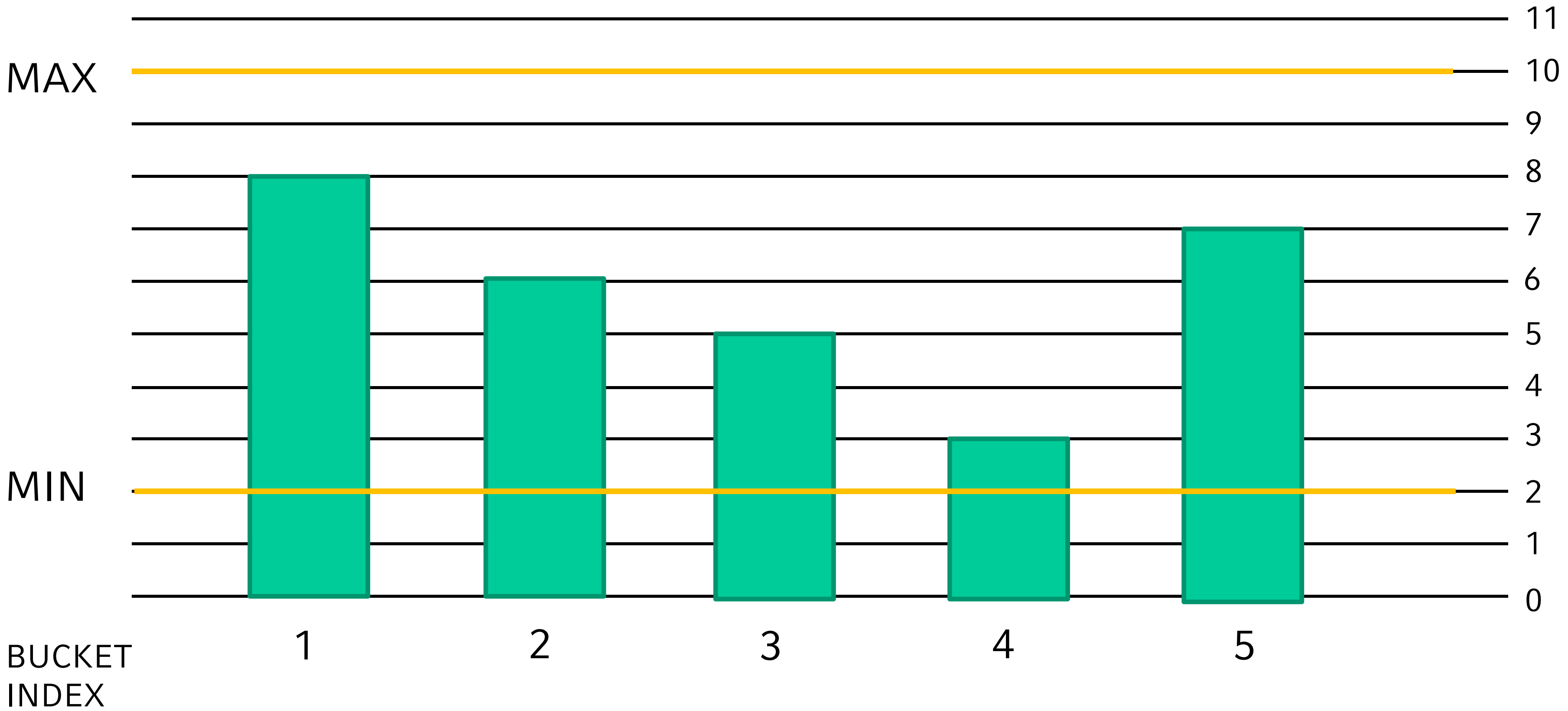
Assignment 9-1

LMU

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

Mode: DELETING

DELETE 1

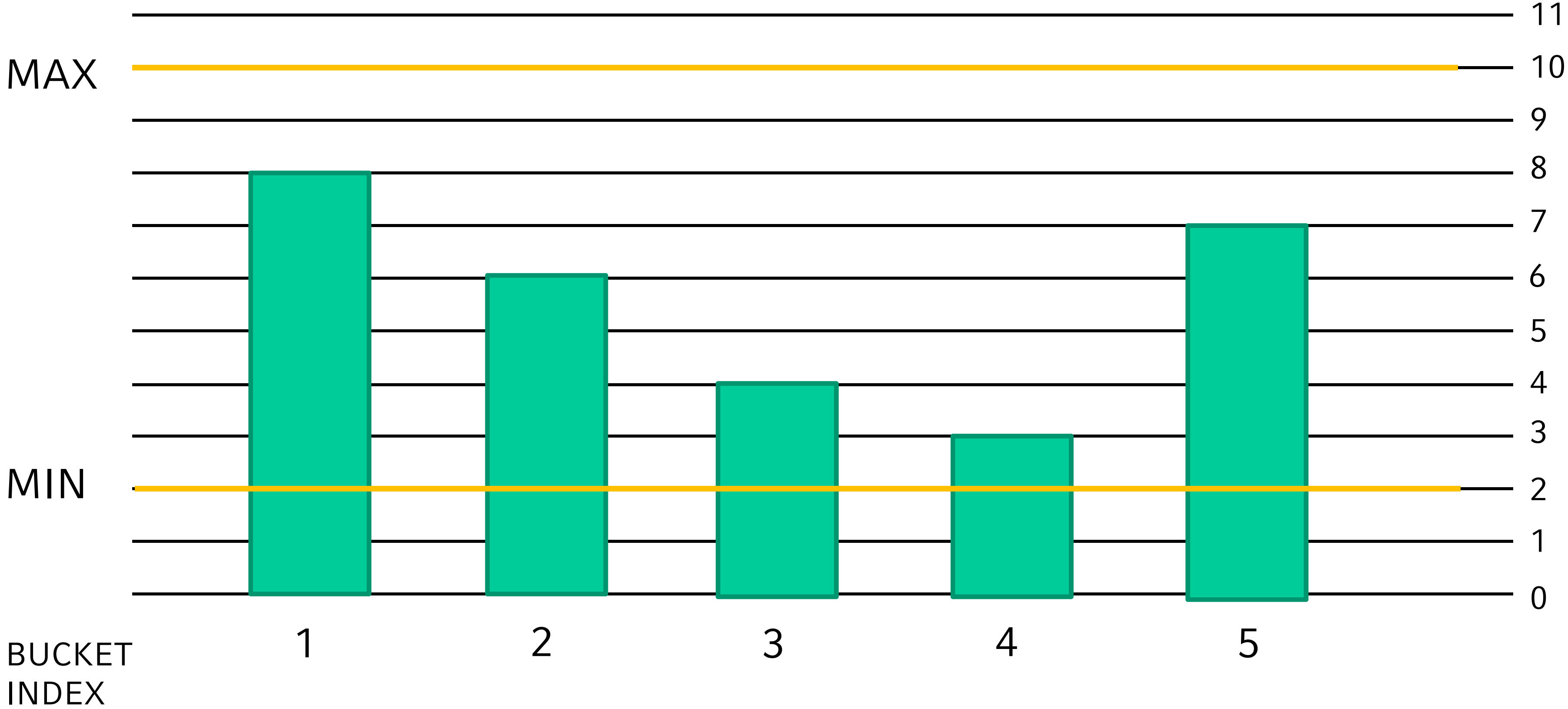


Assignment 9-1

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

Mode: DELETING

DELETE 3

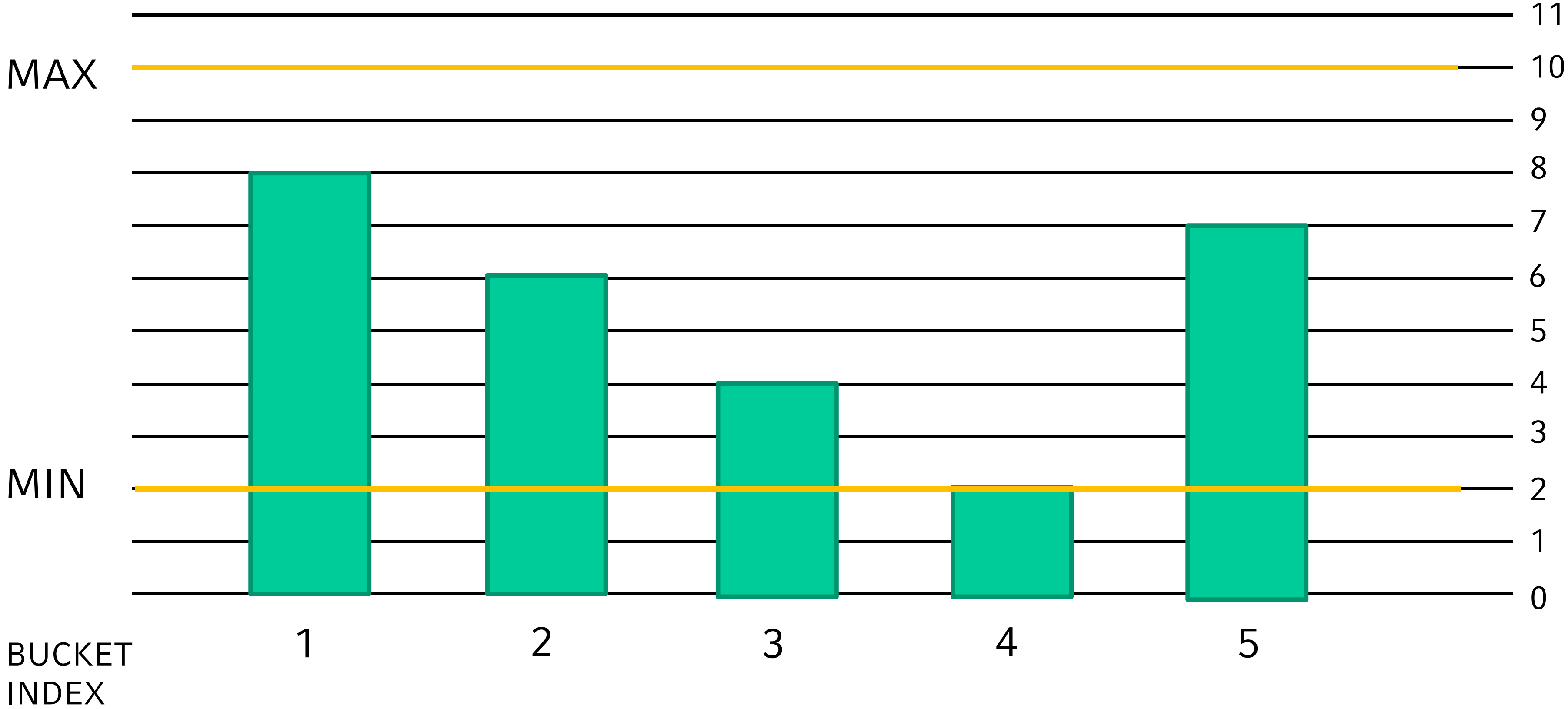


Assignment 9-1

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

Mode: DELETING

DELETE 4

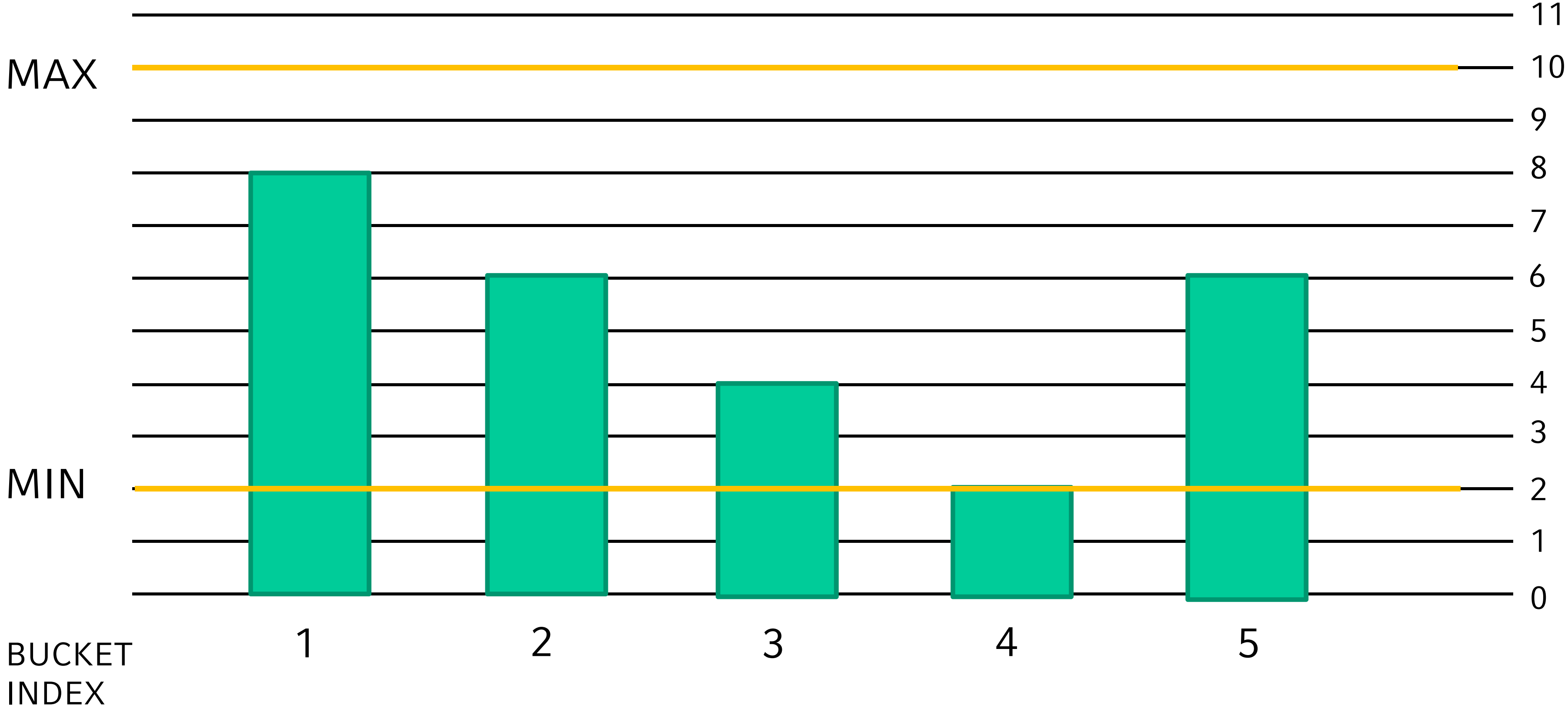


Assignment 9-1

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

Mode: DELETING

DELETE 5

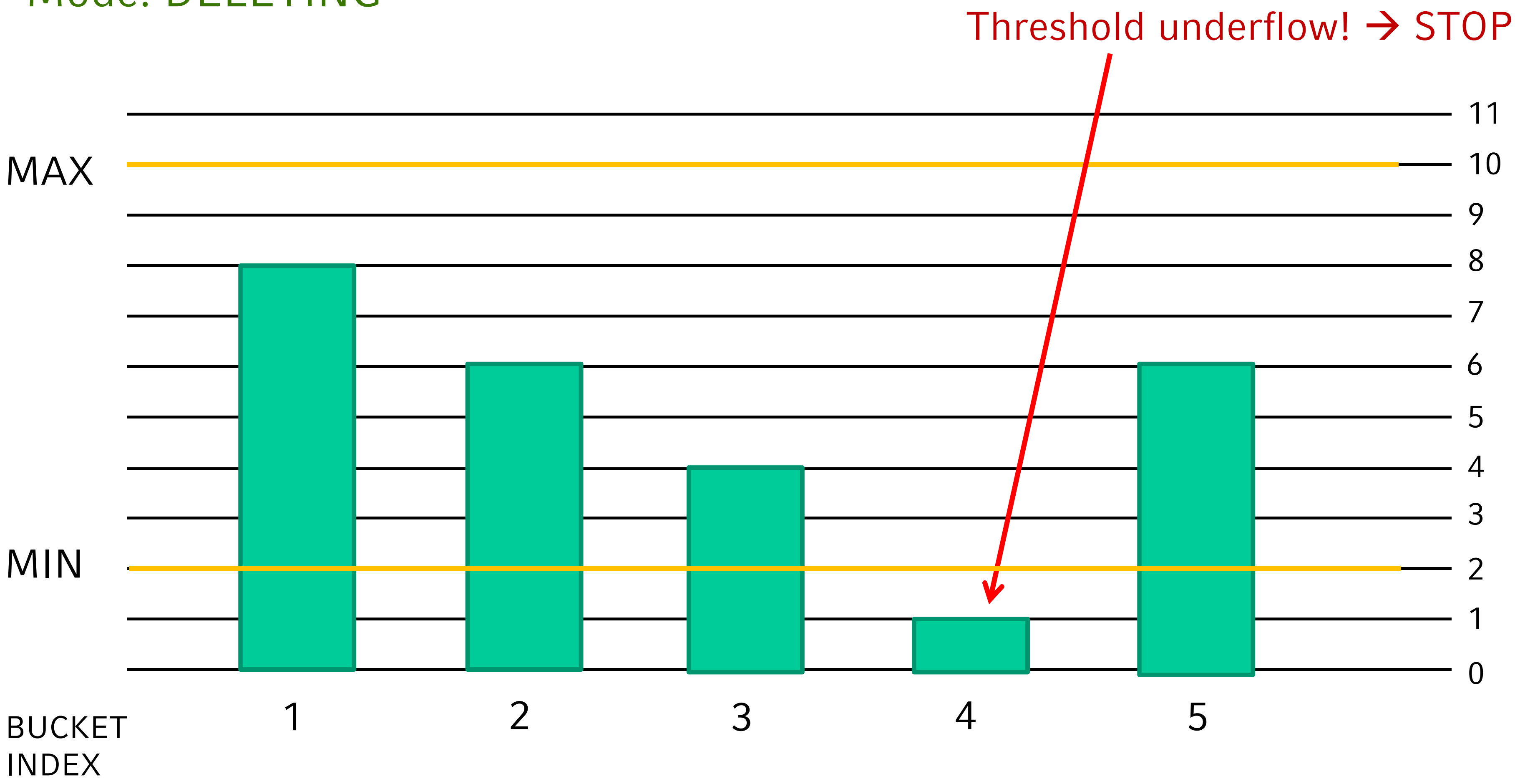


Assignment 9-1

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

Mode: DELETING

DELETE 4



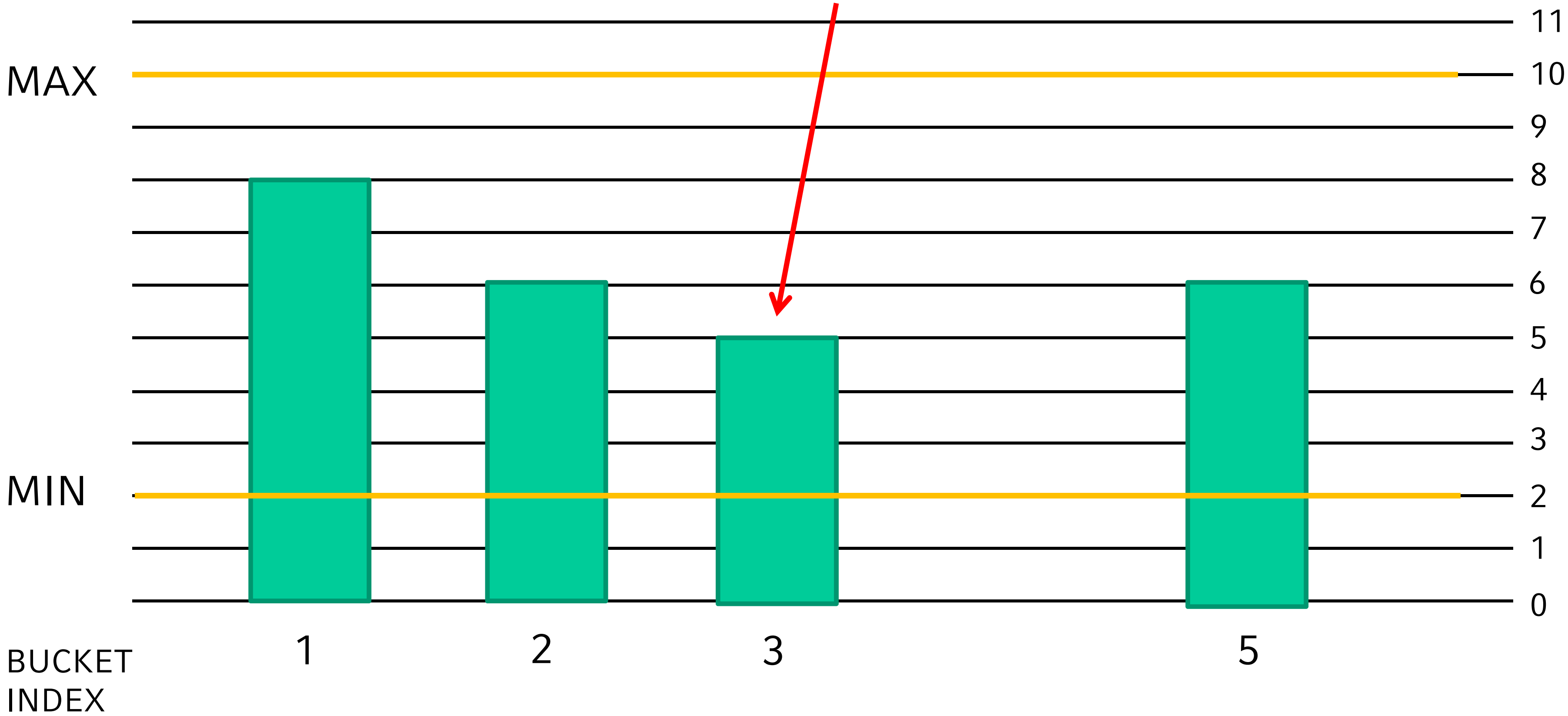
Assignment 9-1

Merge & Split

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

Mode: DELETING

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



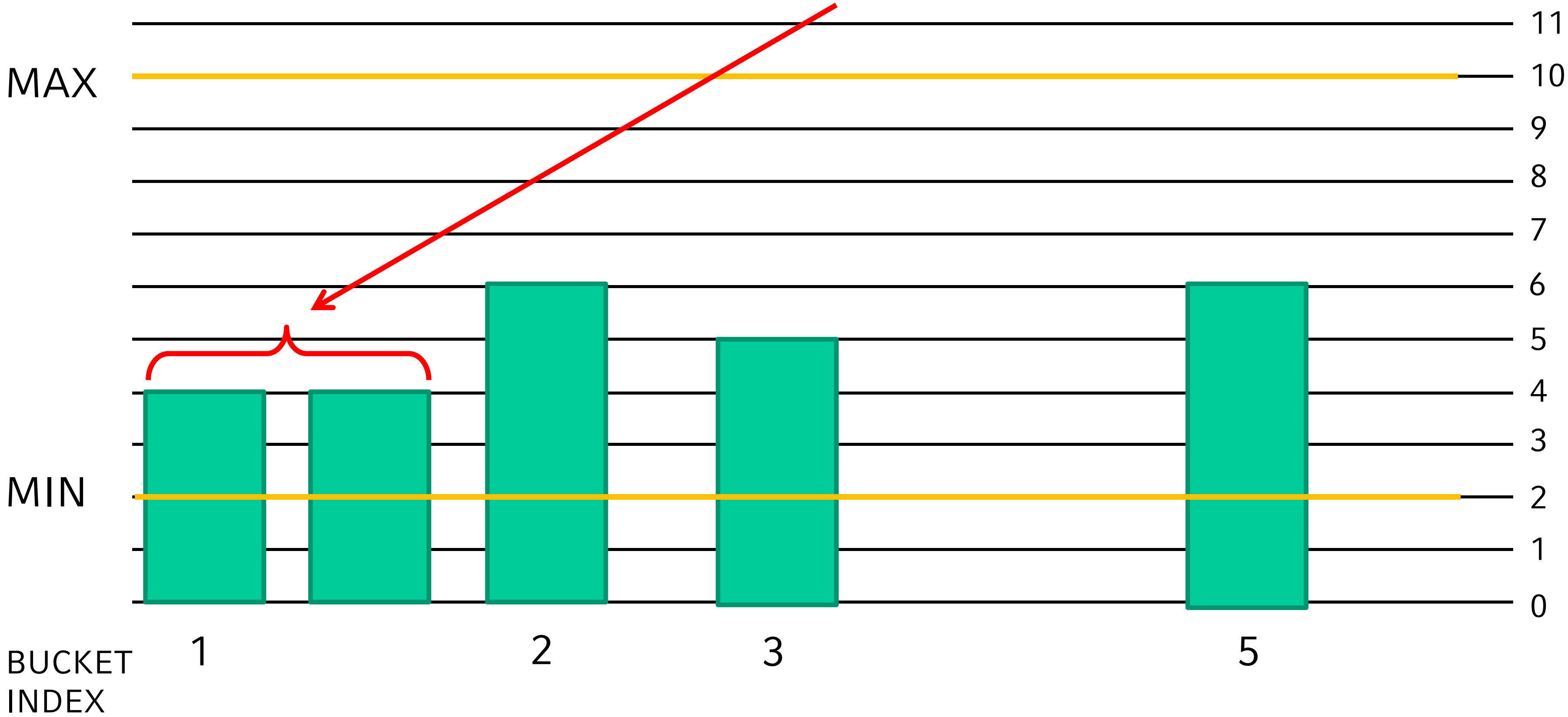
Assignment 9-1

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)



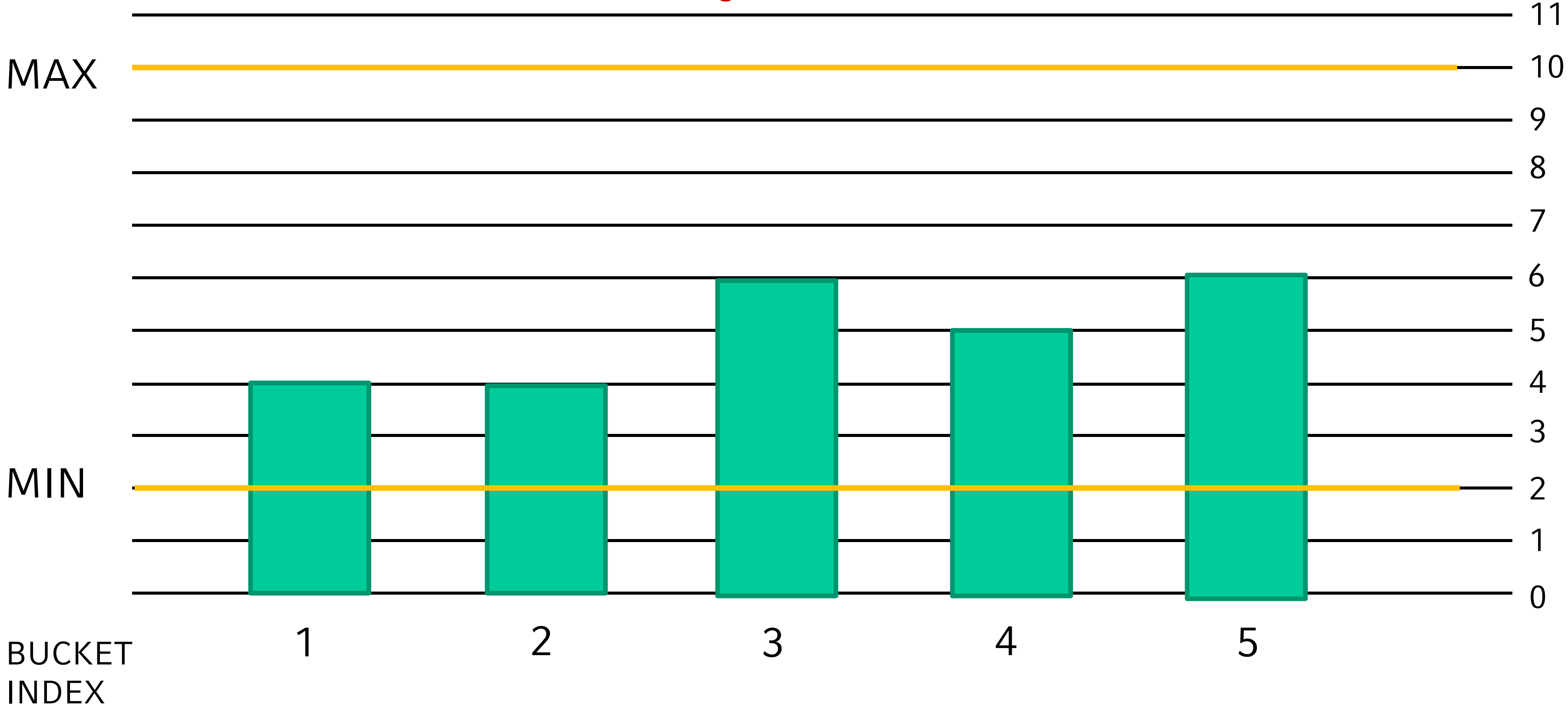
Assignment 9-1

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

Split & Merge

Mode: DELETING

Assign new indices



Assignment 9-2

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

DATABASE SYSTEMS GROUP

Assignment 9-2

LMU

The formula for detecting changes is:

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

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:

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

Assignment 9-2

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
$G_t := 0$

Assignment 9-3

Exponential Histograms

Purpose: solve the problem of counting number of x within a sliding window of size N

Given:

Sequence S = (x, x, o, x, o, o, x, x, x, x, o, x, x, o, x, x)

Window size $N = 8$

Error parameter $\epsilon = \frac{1}{2}$

Assignment 9-3

Sequence S = (x, x, o, x, o, o, x, x, x, x, o, x, x, o, x, x)

Window size $N = 8$

Error parameter $\epsilon = \frac{1}{2}$

Max. # of buckets of same size $\tau = \frac{\left|\frac{1}{\epsilon}\right|}{2} + 2 = 3$

Timest. t_i	Buckets b_i	Element x_i	TOTAL	LAST	$\# buckets$ $of\ same\ size = \tau\,?$
1	1_1	x	1	0	no
2	$1_1, 1_2$	x	2	0	no
3	$1_1, 1_2$	o	2	0	no
4	$1_1, 1_2, 1_4$ $\rightarrow 2_2, 1_4$	x	3	2	yes

Merge two oldest buckets of same size with the largest timestamp of both buckets!

Assignment 9-3

Timest. t_i	Buckets b_i	Element x_i	TOTAL	LAST	# *buckets of same size* $= \tau$?
1	1_1	x	1	0	no
2	1_1, 1_2	x	2	0	no
3	1_1, 1_2	o	2	0	no
4	$2_2, 1_4$	x	3	2	yes
5	$2_2, 1_4$	o	3	2	no
6	$2_2, 1_4$	o	3	2	no
7	$2_2, 1_4$, 1_7	x	4	2	no
8	$2_2, 1_4$, 1_7, 1_8 → $2_2, 2_7$, 1_8	x	5	2	Yes
9	$2_2, 2_7$, 1_8, 1_9	x	6	2	no

Merge two oldest buckets of same size with the largest timestamp of both buckets!

Assignment 9-3

Timest. t_i	Buckets b_i	Element x_i	TOTAL	LAST	# buckets of same size = τ ?
10	$2_2, 2_7, 1_8, 1_9, 1_{10}$ → $2_7, 1_8, 1_9, 1_{10}$ → $2_7, 2_9, 1_{10}$	x	7 7-2=5	2	yes

Merge two oldest buckets of same size with the largest timestamp of both buckets!

3. $b_l := b_{l-1} \rightarrow 2_7$
$\text{LAST} = b_l.size \rightarrow 2$

1. $TOTAL = TOTAL - b_l.size \rightarrow 7 - 2 = 5$

2. Oldest timestamp $t_l \leq t_i - N \rightarrow 2 \leq 10 - 8$
drop the oldest bucket 2_2

Assignment 9-3

Timest. t_i	Buckets b_i	Element x_i	TOTAL	LAST	$\# \; buckets$ $of \; same \; size = \tau \,?$
10	$2_7, 2_9, 1_{10}$	x	5	2	yes
11	$2_7, 2_9, 1_{10}$	o	5	2	no
12	$2_7, 2_9, 1_{10}, 1_{12}$	x	6	2	no
13	$2_7, 2_9, 1_{10}, 1_{12}, 1_{13}$ $\rightarrow 2_7, 2_9, 2_{12}, 1_{13}$ $\rightarrow 4_9, 2_{12}, 1_{13}$	x	7	4	yes

Merge two oldest buckets of same size with the largest timestamp of both buckets!

Last bucket was merged!
$LAST$
$:= size \; of \; the \; new \; created \; last \; bucket$
$= 4$

Merge two oldest buckets of same size with the largest timestamp of both buckets!

Assignment 9-3

Timest. t_i	Buckets b_i	Element x_i	TOTAL	LAST	*# buckets of same size* $= \tau$?
13	$4_9, 2_{12}, 1_{13}$	x	7	4	yes
14	$4_9, 2_{12}, 1_{13}$	o	7	4	no
15	$4_9, 2_{12}, 1_{13}, 1_{15}$	x	8	4	no
16	$4_9, 2_{12}, 1_{13}, 1_{15}, 1_{16}$ $\rightarrow 4_9, 2_{12}, 2_{15}, 1_{16}$	x	9	4	yes

Assignment 9-3

Timest. t_i	Buckets b_i	Element x_i	TOTAL	LAST	# *buckets of same size* $= \tau$?
13	$4_9,\ 2_{12},\ 1_{13}$	x	7	4	yes

Estimating total number of x's within the sliding window of size 8 in the exponential histogram:

x's = EH.TOTAL – EH.LAST/2 = 7 - 4/2 = 5

Sequence S = (x, x, o, x, o, $\boxed{\text{o, x, x, x, x, o, x, x,}}$ o, x, x)

Exact number of x's in sliding window [6:13] : 6

Assignment 9-4

Hoeffding Trees

Core idea: For choosing the best split attribute for a node, a small subset of the training examples may suffice

Question: How many instances are required?

Solution: Utilize the Hoeffding bound

Assignment 9-4

Given:

8 examples of drivers with the attributes:

- Time since getting the driving license (1-2 years, 2-7 years, > 7 years)
- Gender (female, male)
- Residential area (urban, rural)

Further: $\delta = 0.2, \quad N_{min} = 2$

- Use information gain
- Output is norminal risk class → two attributes, R=1

Assignment 9-4

Person	Time since license	Gender	Area	Risk class
1	1-2	m	urban	low
2	2-7	m	rural	high
3	>7	f	rural	low
4	1-2	f	rural	high
5	>7	m	rural	high
6	1-2	m	rural	high
7	2-7	f	urban	low
8	2-7	m	urban	low

Assignment 9-4

$$\boxed{\varepsilon = \sqrt{\frac{R^2 \ln(1/\delta)}{2n}} = \sqrt{\frac{1 \ln(1/\delta)}{2n}} = \sqrt{\ln(1/\delta)} \cdot \sqrt{1/(2n)}}$$

Where

- Confidence $\boldsymbol{\delta}$: what probability do we allow of 'failure'? (How much do we accept a deviation $> \varepsilon$)
- Range **R**: e.g. a probability range from 0 to 1
- # of training examples **n**
- Accuracy $\boldsymbol{\varepsilon}$: How much do we want to allow the empirical mean to differ from the true mean

For $n = 2,4,6,8$ this yields:

$$\varepsilon_2 \approx 0.634$$
$$\varepsilon_4 \approx 0.448$$
$$\varepsilon_6 \approx 0.366$$
$$\varepsilon_8 \approx 0.317$$

Assignment 9-4

RECAP: Entropy and Information Gain (IG)

- T: a set of training objects
- T_i: a partition of T
- A: attribute
- k: # of classes
- c_i: a class
- p_i: a frequency

$$\text{entropy}(T) = \begin{cases} 0,\ if\ p_i = 1\ for\ any\ class\ c_i \\ 1,\ if\ \exists\ k = 2\ classes\ with\ p_i = {}^1/_2\ \ for\ each\ i \\ -\sum_{i=1}^{k} p_i\ \cdot\ log_2 p_i, else \end{cases}$$

$$IG(T, A) = entropy(T)\ - \sum_{i=1}^{m} \frac{|T_i|}{|T|} \cdot entropy(T_i)$$

Assignment 9-4

We initialize an empty tree. Now insert the first two records:

Person	Time since license	Gender	Area	Risk class
1	1-2	m	urban	low
2	2-7	m	rural	high

1. Compute the entropy for the first two records:
   $entropy(D_2) = 1$, due to the fact that both classes have a probability of 50%

2. Compute the information gain (IG) for all three attributes (time, gender, area):
   $IG(time, D_2) = entropy(D_2) - 0.5 \text{ entropy}(D_2|t = 1-2) + 0.5 \text{ entropy } (D_2|t = 2-7) + 0\ entropy(D_2|t > 7) = 1 - (0 + 0 + 0) = 1$

   $IG(gender, D_2) = entropy(D_2) - 1 \text{ entropy}(D_2|g = m) + 0 \text{ entropy } (D_2|g = f) = 1 - (1 + 0) = 0$

   $IG(area, D_2) = entropy(D_2) - 0.5 \text{ entropy}(D_2|a = u) + 0.5 \text{ entropy } (D_2|a = r) = 1 - (0 + 0) = 1$

Assignment 9-4

3. Compare the best with the second best result:
   $IG(time, D_2) - IG(area, D_2) = 1 - 1 = 0 < \varepsilon_2 \approx 0.634$
   → continue with more samples!

Assignment 9-4

Now take two more records (the first four records)

Person	Time since license	Gender	Area	Risk class
1	1-2	m	urban	low
2	2-7	m	rural	high
3	>7	f	rural	low
4	1-2	f	rural	high

Assignment 9-4

Again, proceed as follows:

1. Compute the entropy for the first four records:
   $entropy(D_4) = 1$, due to the fact that both classes have a probability of 50%

2. Compute the information gain (IG) for all three attributes (time, gender, area):
   $IG(time, D_4) = entropy(D_4) - \frac{2}{4}\,\text{entropy}(D_4|t = 1-2) + \frac{1}{4}\,\text{entropy}\,(D_4|t = 2-7) + \frac{1}{4}\,entropy(D_4|t > 7) = 1 - \left(\frac{2}{4} * 1 + 0 + 0\right) = 0.5$

   $IG(gender, D_4) = entropy(D_4) - \frac{2}{4}\,\text{entropy}(D_4|g = m) + \frac{2}{4}\,\text{entropy}\,(D_4|g = f) = 1 - \left(\frac{1}{2} * 1 + \frac{1}{2} * 1\right) = 0$

   $IG(area, D_4) = entropy(D_4) - \frac{1}{4}\text{entropy}(D_4|a = u) + \frac{3}{4}\,\text{entropy}\,(D_4|a = r) \approx 1 - \left(0 + \frac{3}{4} * 0.637\right) \approx 0.523$

Assignment 9-4

3. Compare the best with the second best result:
   $IG(time, D_4) - IG(area, D_4) = 0.523 - 0.5 \approx 0.023 < \varepsilon_4 \approx 0.448$
   → continue with more samples!

Assignment 9-4

Now take two more records (the first six records)

Person	Time since license	Gender	Area	Risk class
1	1-2	m	urban	low
2	2-7	m	rural	high
3	>7	f	rural	low
4	1-2	f	rural	high
5	>7	m	rural	high
6	1-2	m	rural	high

Assignment 9-4

Again, proceed as follows:

1. Compute the entropy for the first six records:

$$entropy(D_6) = -\frac{2}{3}\log\left(\frac{2}{3}\right) - \frac{1}{3}\log\left(\frac{1}{3}\right) \approx 0.637$$

2. Compute the information gain (IG) for all three attributes (time, gender, area):

$IG(time, D_6) = entropy(D_6) - \frac{3}{6}\text{entropy}(D_6|t = 1 - 2) + \frac{1}{6}\text{ entropy }(D_6|t = 2 - 7) + \frac{2}{6}\ entropy(D_6|t > 7) \approx 0.637 - \left(\frac{3}{6} * 0.637 + 0 + \frac{2}{6} * 1\right) \approx -0.015$

$IG(gender, D_6) = entropy(D_6) - \frac{4}{6}\text{entropy}(D_6|g = m) + \frac{2}{6}\text{ entropy }(D_6|g = f) \approx 0.637 - \left(\frac{4}{6} * 0.562 + \frac{2}{6} * 1\right) \approx -0.072$

$IG(area, D_6) = entropy(D_6) - \frac{1}{6}\text{entropy}(D_6|a = u) + \frac{5}{6}\text{entropy }(D_6|a = r) \approx 0.637 - \left(0 + \frac{5}{6} * 0.5004\right) \approx 0.220$

Assignment 9-4

3. Compare the best with the second best result:
   $IG(area, D_6) - IG(time, D_6) = 0.220 - -0.015 \approx 0.235 < \varepsilon_6 \approx 0.366$
   → continue with more samples!

Assignment 9-4

Now take two more records (all eight records)

Person	Time since license	Gender	Area	Risk class
1	1-2	m	urban	low
2	2-7	m	rural	high
3	>7	f	rural	low
4	1-2	f	rural	high
5	>7	m	rural	high
6	1-2	m	rural	high
7	2-7	f	urban	low
8	2-7	m	urban	low

Assignment 9-4

Again, proceed as follows:

1. Compute the entropy for all eight records:

$$entropy(D_8) = 1$$

2. Compute the information gain (IG) for all three attributes (time, gender, area):

$$IG(time, D_8) = entropy(D_8) - \frac{3}{8}\text{entropy}(D_8|t = 1-2) + \frac{3}{8}\text{ entropy }(D_8|t = 2-7) + \frac{2}{8}entropy(D_8|t > 7) \approx 1 - \left(\frac{3}{8} * 0.637 + \frac{3}{8} * 0.637 + \frac{2}{8} * 1\right) \approx 0.273$$

$$IG(gender, D_8) = entropy(D_8) - \frac{5}{8}\text{entropy}(D_8|g = m) + \frac{3}{8}\text{ entropy }(D_8|g = f) \approx 1 - \left(\frac{5}{8} * 0.673 + \frac{3}{8} * 0.637\right) \approx 0.341$$

$$IG(area, D_8) = entropy(D_8) - \frac{3}{8}\text{entropy}(D_8|a = u) + \frac{5}{8}\text{entropy }(D_8|a = r) \approx 1 - \left(0 + \frac{5}{8} * 0.5004\right) \approx 0.687$$

Assignment 9-4

3. Compare the best with the second best result:
   $IG(area, D_8) - IG(gender, D_8) = 0.687 - 0.341 \approx 0.347 > \varepsilon_8 \approx 0.317$
   → split at 'area' attribute!

   → New leafs are empty and have no 'area' attribute.
   → Further splits are not required until new data arrives.

DATABASE SYSTEMS GROUP

Assignment 9-4

Computing the value of δ at which the tree would still consist only of the leaf:

The minimal ε for which a further split would be required: 0.347

$$\varepsilon = \sqrt{\frac{R^2 \ln(1/\delta)}{2n}} \Rightarrow 2n\varepsilon^2 = \ln(1/\delta) \Rightarrow \delta = \frac{1}{\exp(2n\varepsilon^2)} \approx \frac{1}{\exp(16*0.347^2)} \approx 0.1456$$