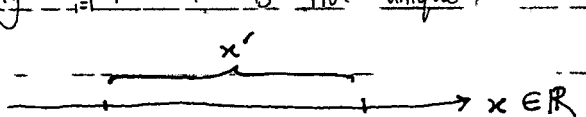
7.11 Example where x minimizing $\sum_{i=1}^{n} |a_i - x|$ is not unique:



x' can be any number within $[-1, 1]$ and they all minimise $\sum_{i=1}^{n} |a_i - x|$

where $a_1 = -1$, $a_2 = 1$

Example where centroid is different from x that minimises $\sum_{i=1}^{n} |a_i - x|$:

Data points are: $-2, -1, 1, 2, 100$

Centroid, $\mu = \frac{-2-1+1+2+100}{5} = 20$

~~$\operatorname{argmin}_x \sum_{i=1}^{n} |a_i - x|$~~ $= 1$ ~~(the median)~~

The points 1 and 20 are quite far apart.

7.12 Want to show that $\frac{1}{n^2} \sum_{i,j} a_i a_j^T = \frac{1}{n} \sum_{i=1}^{n} a_i C^T$

Proof: We have $C = \frac{1}{n} \sum_{i=1}^{n} a_i$

$$\text{RHS} = \frac{1}{n} \sum_{i=1}^{n} a_i \left(\frac{1}{n} \sum_{j=1}^{n} a_j^T \right)$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} a_i \sum_{j=1}^{n} a_j^T$$

$$= \frac{1}{n^2} \sum_{i=1}^{n} \sum_{j=1}^{n} a_i a_j^T$$

$$= \text{L.H.S.}$$

Hence average cluster similarity is the same as computing the average similarity of each point with the centroid of the cluster. ∎

3.3.3 (a)

Element (row=r)	S_1	S_2	S_3	S_4	$h_1(r)$	$h_2(r)$	$h_3(r)$
0	0	1	0	1	1	2	2
1	0	1	0	0	3	5	1
2	1	0	0	1	5	2	0
3	0	0	1	0	1	5	5
4	0	0	1	1	3	2	4
5	1	0	0	0	5	5	3

computing min hash signatures: Init:

S_1	S_2	S_3	S_4
∞	∞	∞	∞
∞	∞	∞	∞
∞	∞	∞	∞

row 0 :

S_1	S_2	S_3	S_4
∞	1	∞	1
∞	2	∞	2
∞	2	∞	2

row 1 :

S_1	S_2	S_3	S_4
∞	1	∞	1
∞	2	∞	2
∞	1	∞	2

row 2:

S_1	S_2	S_3	S_4
5	1	∞	0
2	2	∞	2
0	1	∞	0

1 (since 1 < 5, keep unchanged)
Refers to: MMD p. 84
[set SIG(i, c) to the smaller of the current value of SIG(i, c) and h i (r).]

row 3:

S_1	S_2	S_3	S_4
5	1	1	0 1
2	2	5	2
0	1	5	0

row 4:

S_1	S_2	S_3	S_4
5	1	1	0 1
2	2	2	2
0	1	4	0

row 5:

S_1	S_2	S_3	S_4
5	1	1	0 1
2 5	2	2	2
0	1	4	0

My answer:
5 1 1 1
2 2 2 2
0 1 4 0

↓

Final minhash signatures

(b) Only $h_3(x) = 5x+2 \pmod 6$ is a true permutation.

3.3.3 (c)

Pair	True JS	Est. JS from minhash sig	
S_1 S_2	0	1/3	0
S_1 S_3	0	1/3	0
S_1 S_4	1/4	2/3	1/3
S_2 S_3	0		2/3
S_3 S_4	1/4	2/3	1/3
S_2 S_4	1/4	2/3	1/3

The estimated JS are mostly accurate, except for the pairs S_2, S_3. This could be due to the small number of minhash signatures, or to the hash collisions.