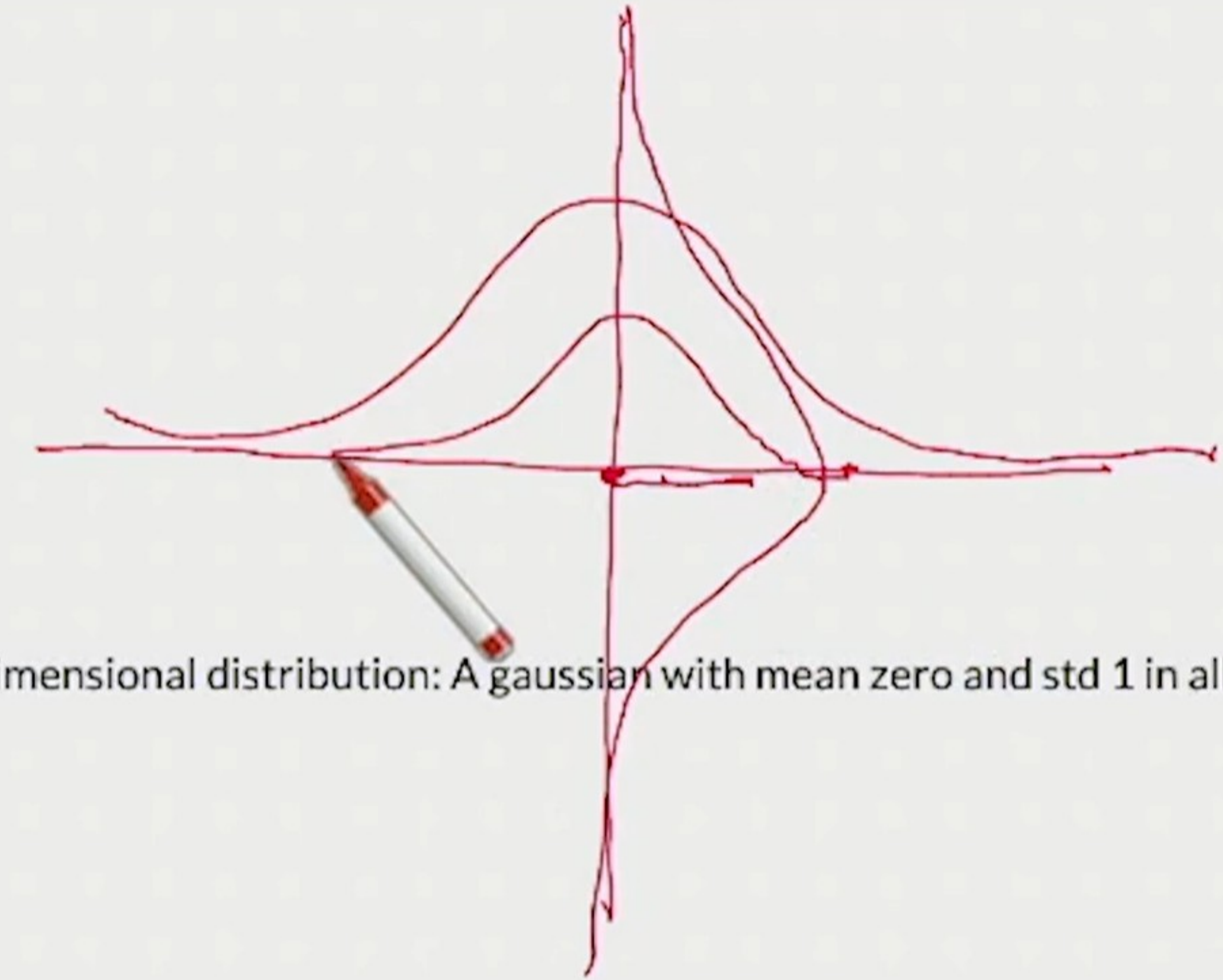
10.2 Data in High Dimensions

DSC 232R, Class 10: RMS

1 Data in High Dimensions/ The Curse of Dimensionality

- Random vectors in high dimension behave very differently from low dimensional vectors
- Our intuition fails us
- For a better understanding, rely on the law of large numbers and central limit theorem

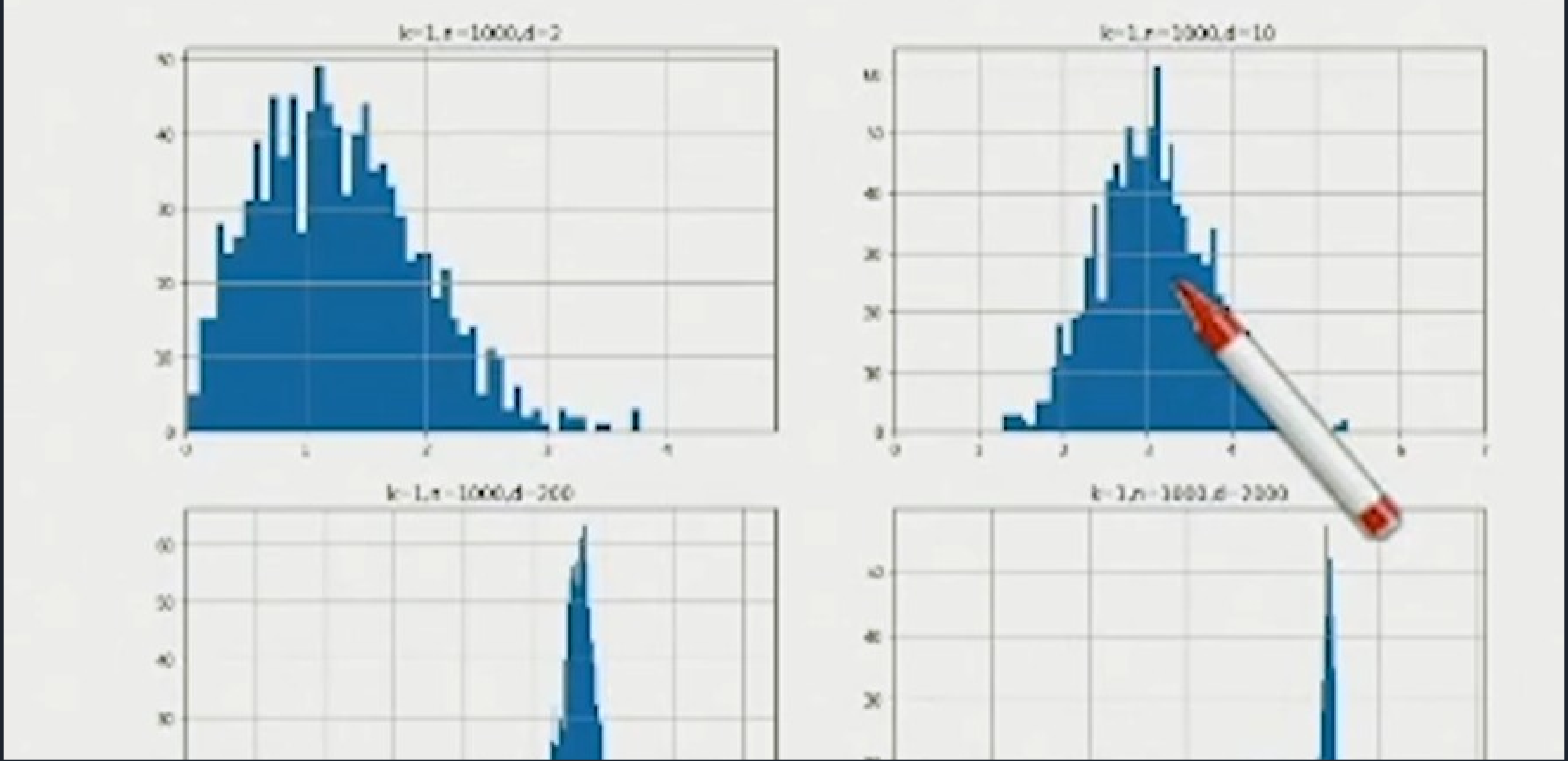


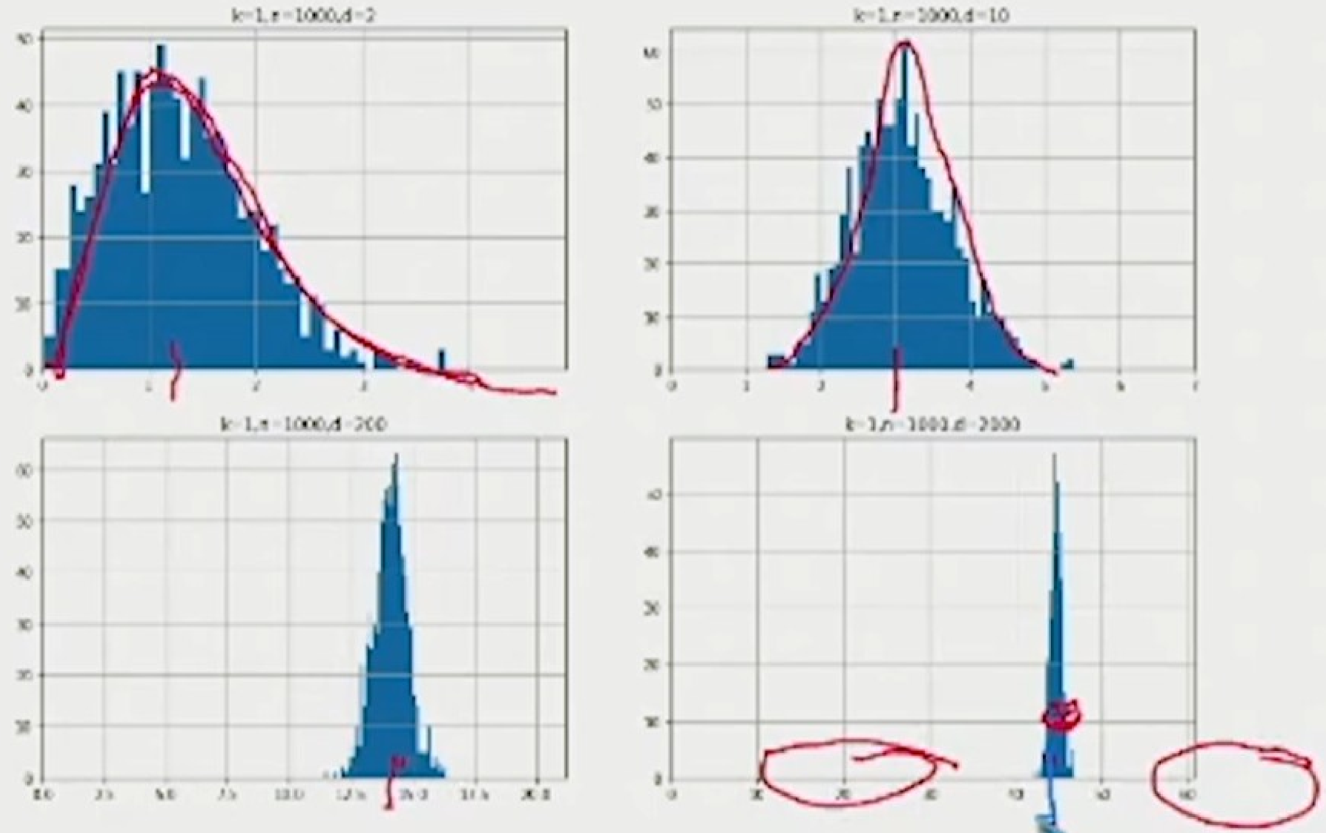
2 Single gaussian

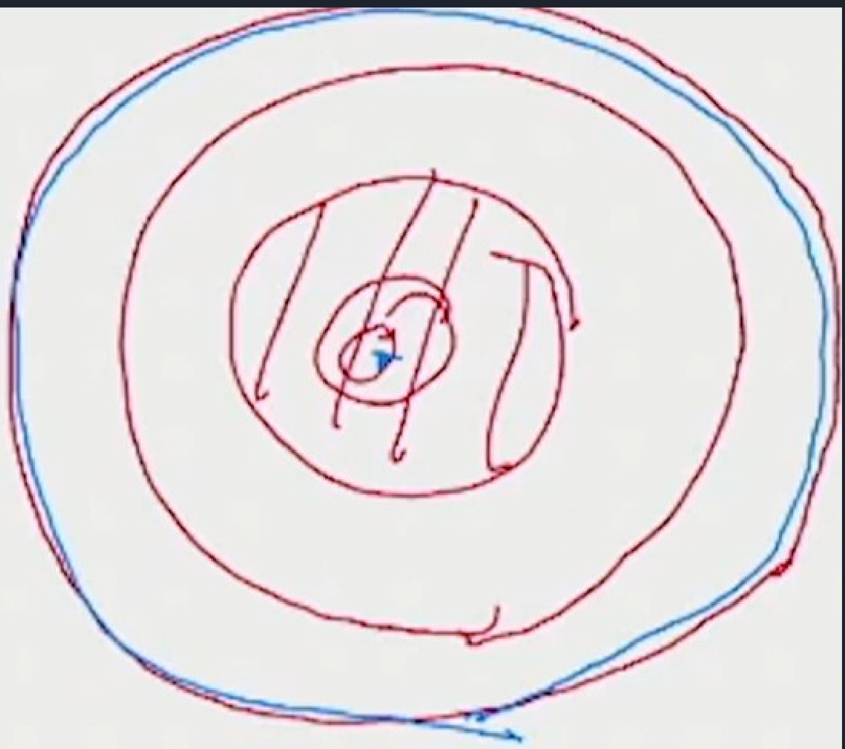
We start with the simplest high dimensional distribution: A gaussian with mean zero and std 1 in all directions.

2.1 The Length of a Random Vector

- $\vec{x}$ is a random vector drawn from a gaussian centered at zero
- For $d = 2$: $||\vec{x}||$ is around std=1
- For $d \to \infty$ the diatribution of $||\vec{x}||$ becomes concentrated around $\sqrt{d}$


k=1, n=1000, d=2
50
40
30
20
10
0
0
1
2
3
4
k=1, n=1000, d=10
60
50
40
30
20
10
0
0
1
2
3
4
6
7
k=1, n=1000, d=200
60
50
40
30
k=1, n=1000, d=2000

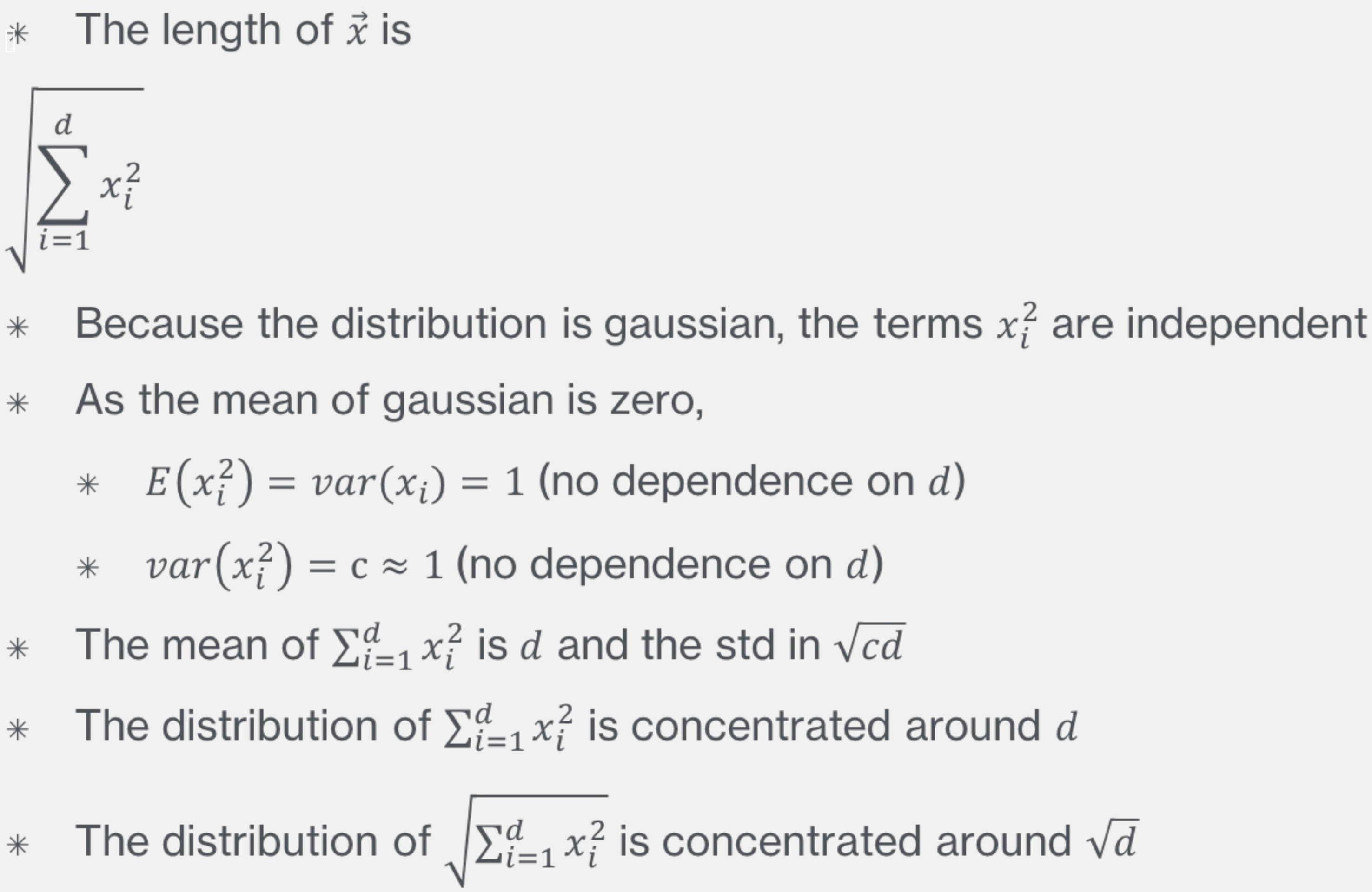

k=1,n=1000,d=2
0
10
20
30
40
50
0
1
2
3
k=1,n=1000,d=10
0
10
20
30
40
50
60
0
1
2
3
4
5
6
7
k=1,n=1000,d=200
0
10
20
30
40
50
60
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
k=1,n=1000,d=2000
0
10
20
30
40
50
0
10
20
30
40
50
60



45

2.2 What is the Explanation?

$$\frac{c\sqrt{d}}{d} = \frac{c}{\sqrt{d}}$$



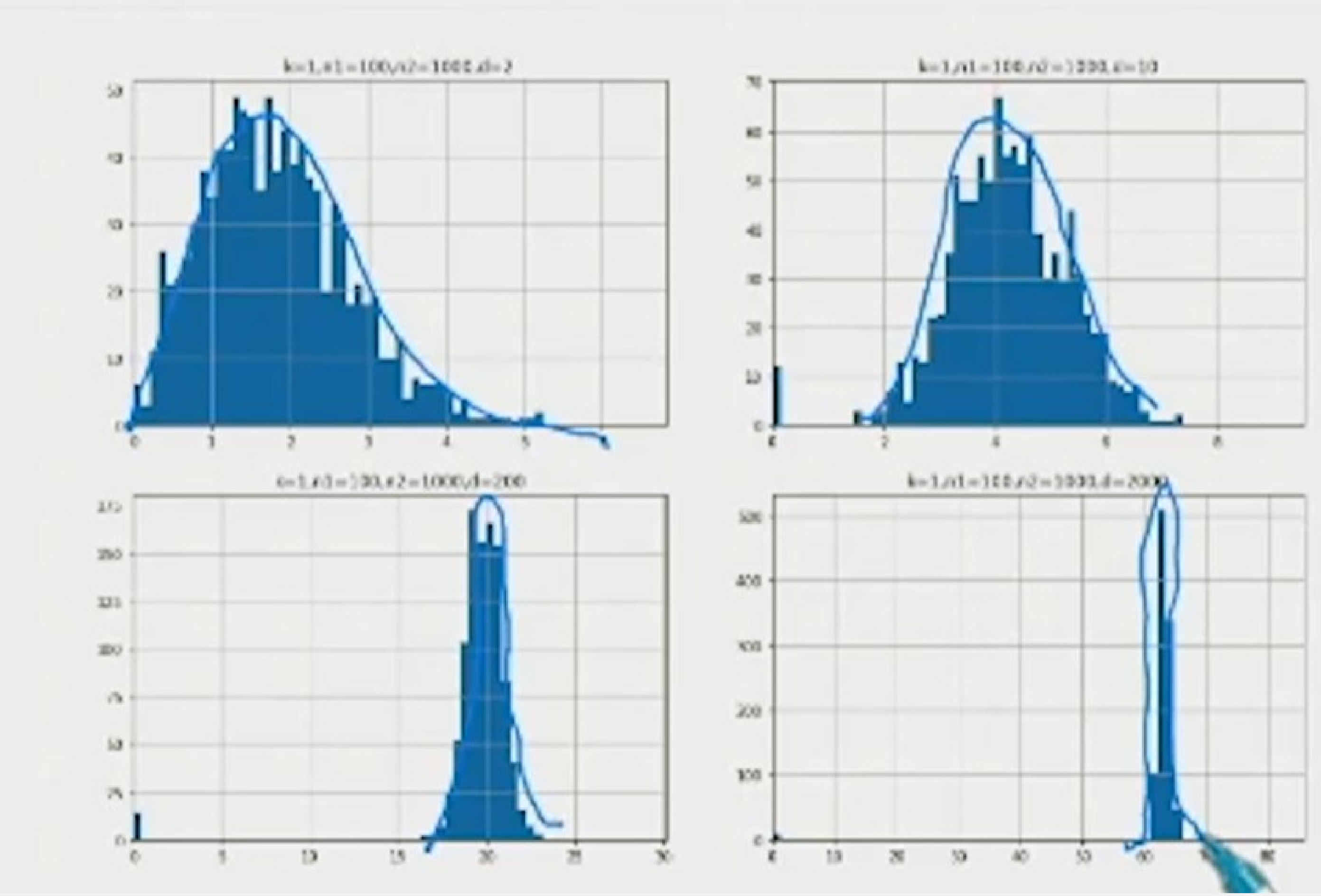
- The length of $\vec{x}$ is

$$\sqrt{\sum_{i=1}^{d} x_i^2}$$

- Because the distribution is gaussian, the terms x_i^2 are independent
- As the mean of gaussian is zero,
  - $E\left(x_i^2\right) = var(x_i) = 1$ (no dependence on d)
  - $var\left(x_i^2\right) = \mathrm{c} \approx 1$ (no dependence on d)
- The mean of $\sum_{i=1}^{d} x_i^2$ is d and the std in $\sqrt{cd}$
- The distribution of $\sum_{i=1}^{d} x_i^2$ is concentrated around d
- The distribution of $\sqrt{\sum_{i=1}^{d} x_i^2}$ is concentrated around $\sqrt{d}$

2.3 What about the Distance Between Two Random Vectors

Drawn from the same gaussian

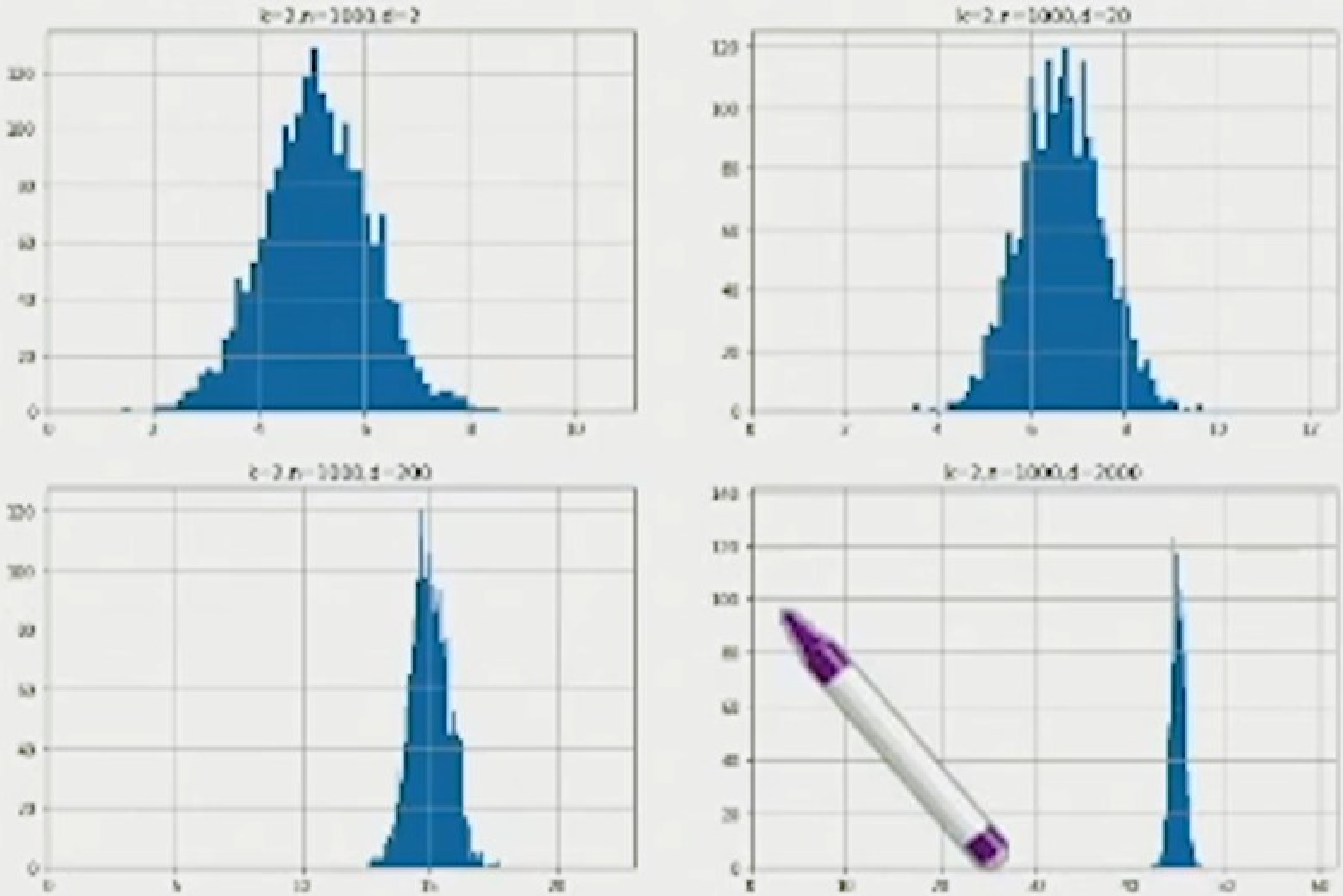


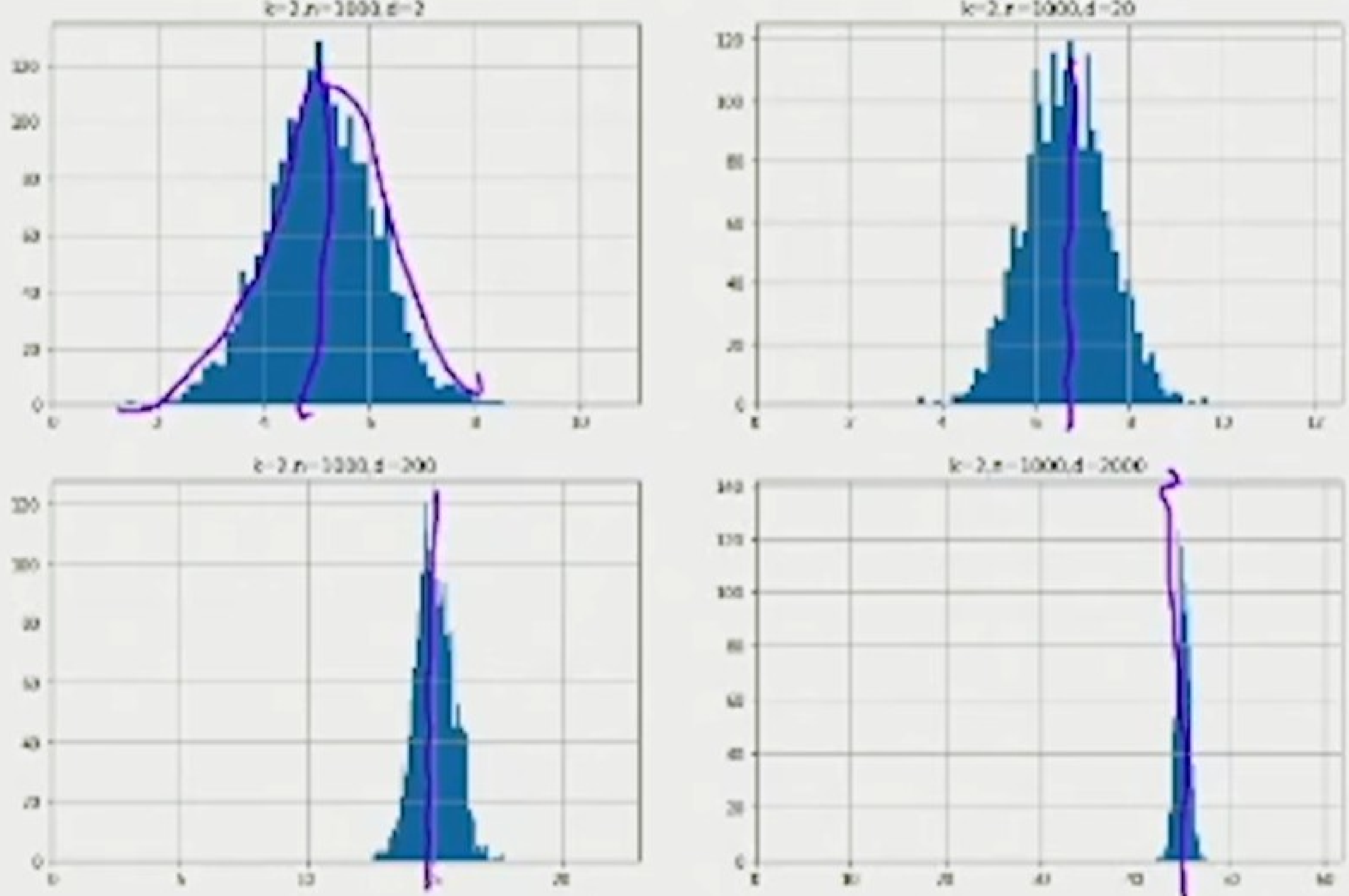
3 Two Gaussians

- Maybe this only happens when the vectors come from a single gaussian?
- We consider data generated from two gaussian distributions
  - mean1 at $x = (-5,0,0,0, \dots 0)$
  - mean2 at $x = (+5,0,0,0, \dots 0)$ All the std are 1

3.1 The Length of a Random Vector

- If the data is two dimensional, it behaves like we would expect – most vectors have length close to 5
- As the dimension increases, the length of the vectors becomes dominated by the coordinates with mean 0


k=2, n=1000, d=2
0
20
40
60
80
100
120
0
2
4
6
8
10
k=2, n=1000, d=20
0
20
40
60
80
100
120
0
2
4
6
8
10
12
k=2, n=1000, d=200
0
20
40
60
80
100
120
0
5
10
15
20
k=2, n=1000, d=2000
0
20
40
60
80
100
120
140
0
10
20
30
40
50
60

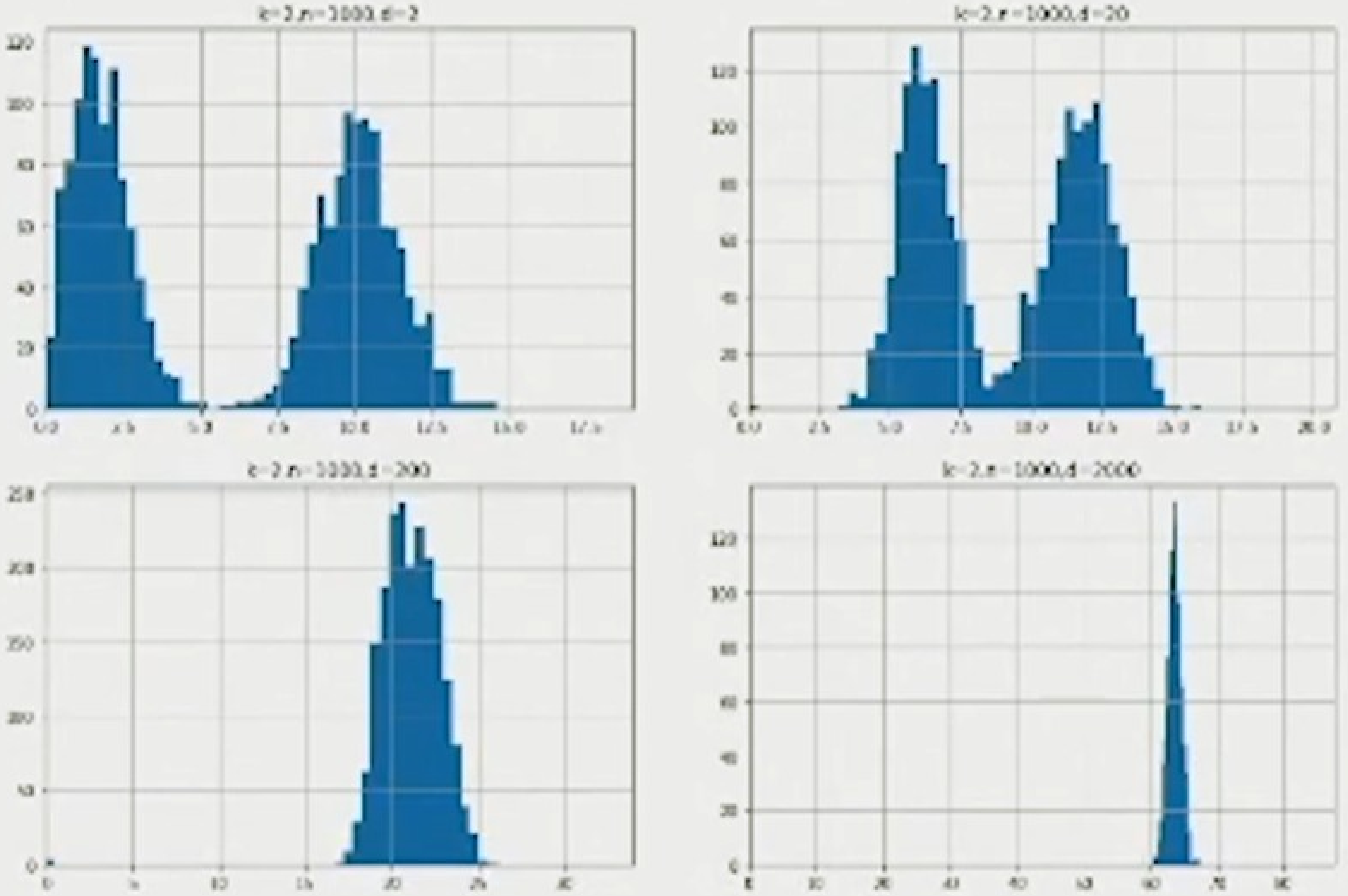

k=2, n=1000, d=2
0
20
40
60
80
100
120
0
2
4
6
8
10
k=2, n=1000, d=20
0
20
40
60
80
100
120
0
2
4
6
8
10
12
k=2, n=1000, d=200
0
20
40
60
80
100
120
0
5
10
15
20
k=2, n=1000, d=2000
0
20
40
60
80
100
120
140
0
10
20
30
40
50
60

3.2 Compare with a Single Gaussian

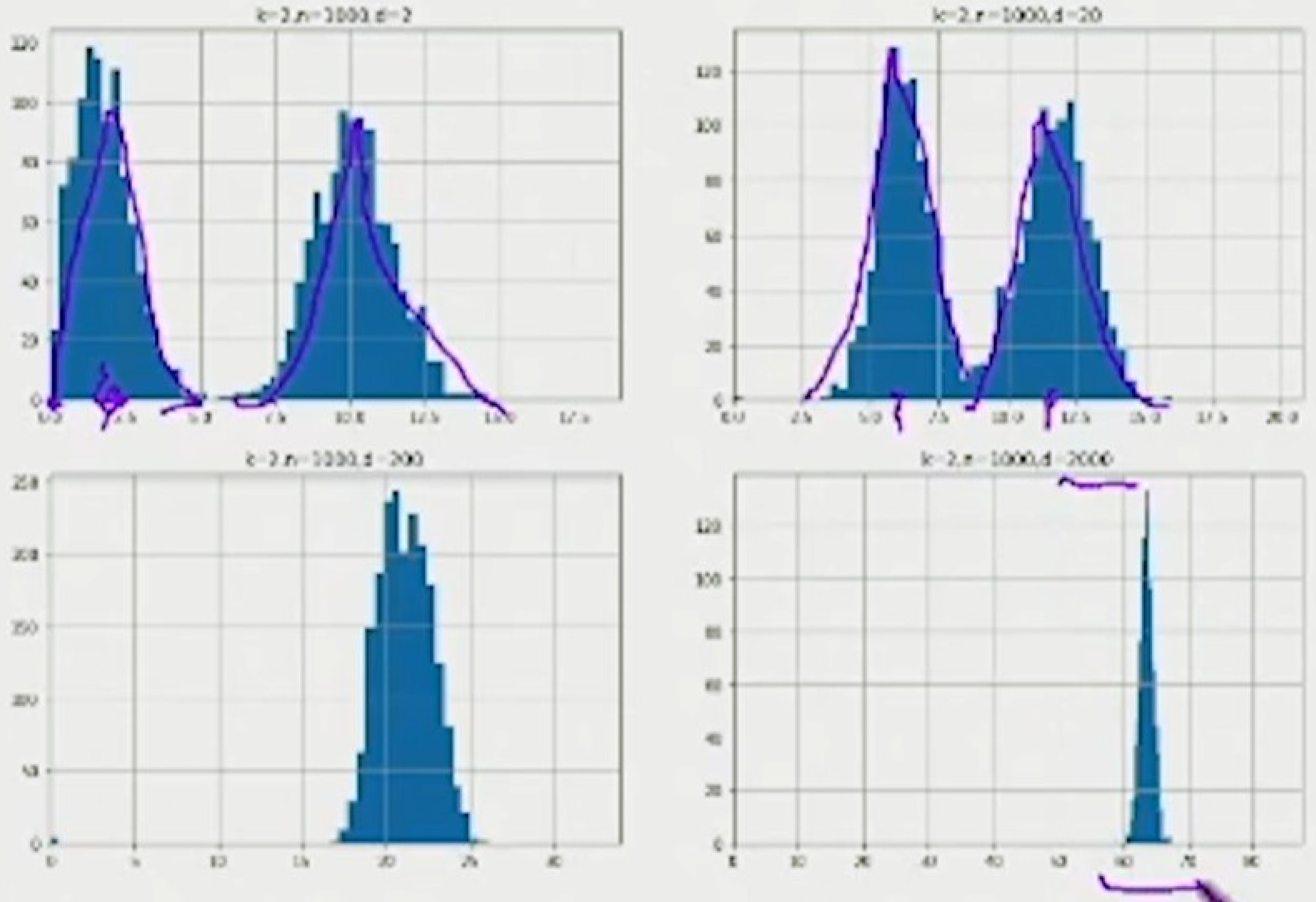
- In $d = 2000$, the distributions are almost the same: concentrated around $\sqrt{2000 \approx 45}$
- We cannot use distance distribution to distinguish $k = 1$ and $k = 2$

3.3 Distance Between Two Points

- We have two spherical gaussians with std 1,
  - mean1 at $x = (-5,0,0,0, \dots 0)$
  - mean2 at $x = (+5,0,0,0, \dots 0)$
- When we have just two dimensions there is a clear distinction between small distances – where the two points are in the same cluster, and large distances, where they belong to a different clusters



k=2,n=1000,d=2
0
20
40
60
80
100
120
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
k=2,n=1000,d=20
0
20
40
60
80
100
120
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
k=2,n=1000,d=200
0
50
100
150
200
250
0
5
10
15
20
25
30
k=2,n=1000,d=2000
0
20
40
60
80
100
120
0
10
20
30
40
50
60
70
80



k=2,n=1000,d=2
0
20
40
60
80
100
120
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
k=2,n=1000,d=20
0
20
40
60
80
100
120
0.0
2.5
5.0
7.5
10.0
12.5
15.0
17.5
20.0
k=2,n=1000,d=200
0
50
100
150
200
250
0
5
10
15
20
25
30
k=2,n=1000,d=2000
0
20
40
60
80
100
120
0
10
20
30
40
50
60
70
80

3.4 K-nearest-neighbors Fails in High Dimensions

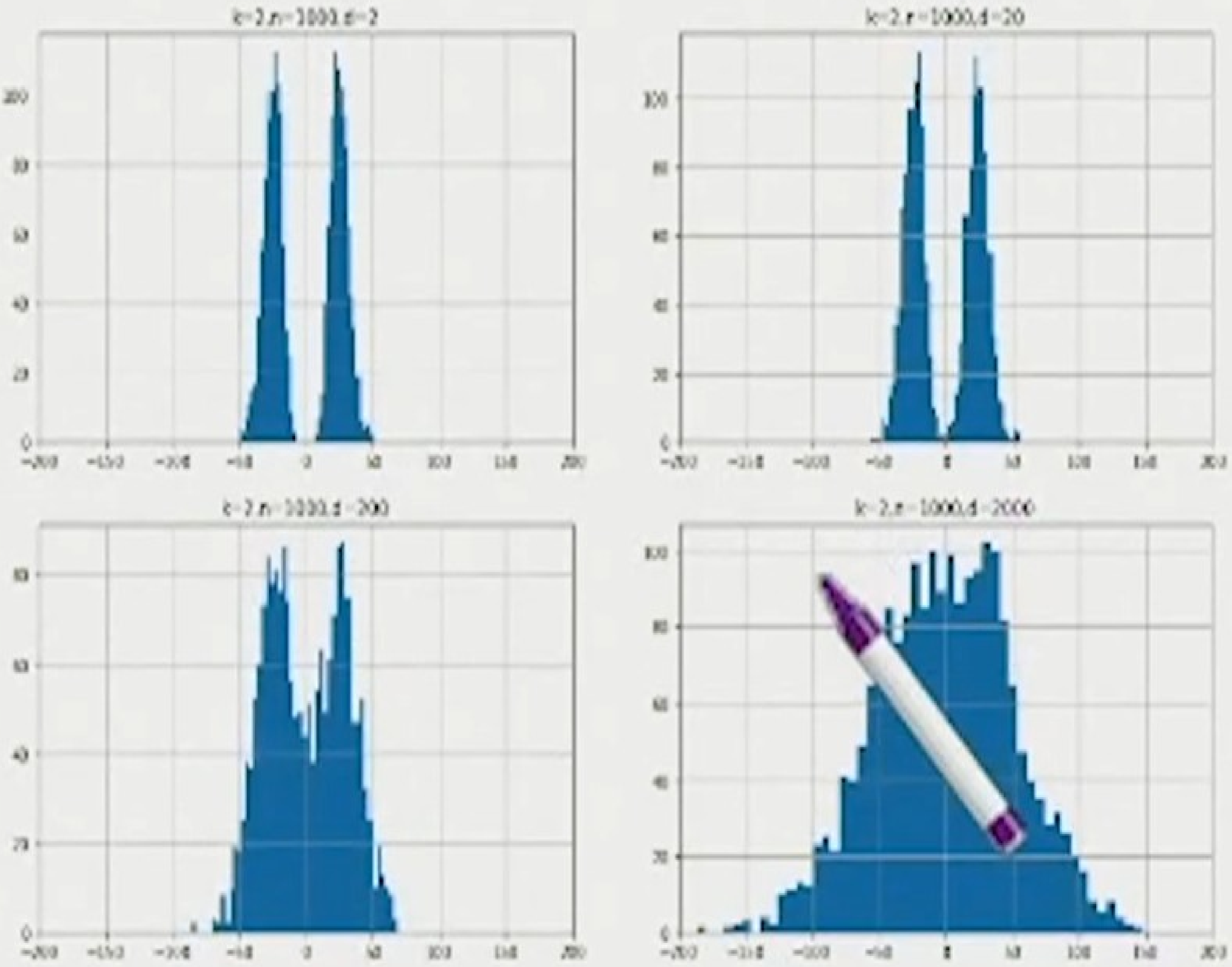
- Suppose one gaussian is labeled +1 and the other -1
- Can we use the k-NN classifier to identify the cluster from which the data point came?
- In low dimension – yes
- In high dimensions – no
- Given two vectors, the distance between them is about $\sqrt{d}$ regardless of whether they are from the same or opposite clusters

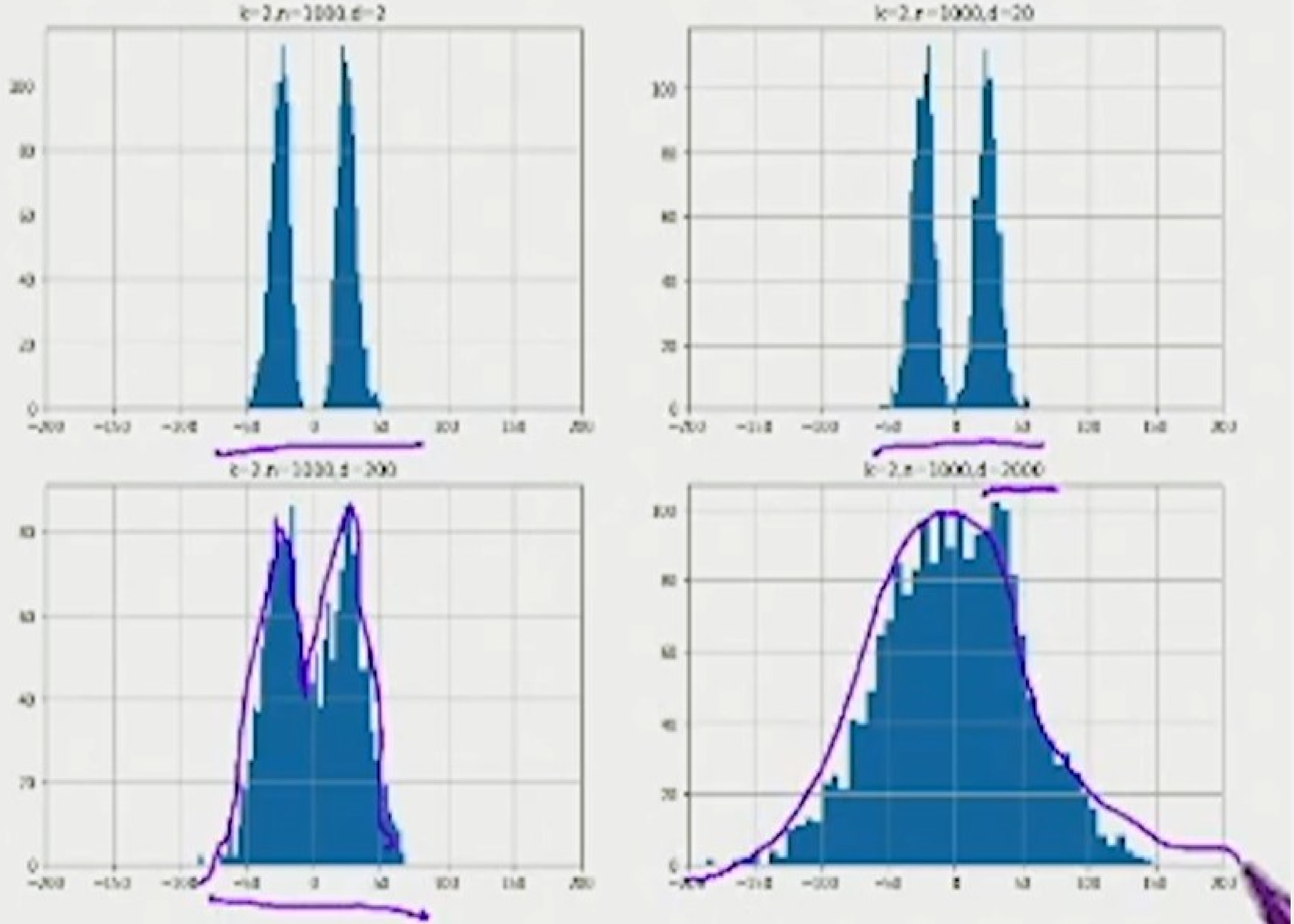
4 Dot Products

- When we take the dot product between two random vectors we expect that, because of the two clusters, the result would be concentrater around

$$5 \times -5 = -25 \;\; \mathit{or} \; 5 \times 5 = +25$$

- That is true for dimension 2 and for dimension 20. But when we reach dimension 200, the peaks have a large overlap, and when we reach 2000 they are indistinguishable


k=2,n=1000,d=2
k=2,n=1000,d=20
k=2,n=1000,d=200
k=2,n=1000,d=2000
0
20
40
60
80
100
−20.0
−15.0
−10.0
−5.0
0
5.0
10.0
15.0
20.0


k=2,n=1000,d=2
k=2,n=1000,d=20
k=2,n=1000,d=200
k=2,n=1000,d=2000
0
20
40
60
80
100
−200
−150
−100
−50
0
50
100
150
200

4.1 Summary

- The 2d and 3D intuitions break down at high dimensions. The lengths of vectors, the distance between vectors and the dot products between vectors all become highly concentrated and therefor not informative
- **The curse of dimensionality**: Most statistical methods break down at high dimensions
- A way out: some high dimensional data has low **intrinsic dimension**
- For example: **small number of PCA eigenvectors explain alarge fraction of the variance**