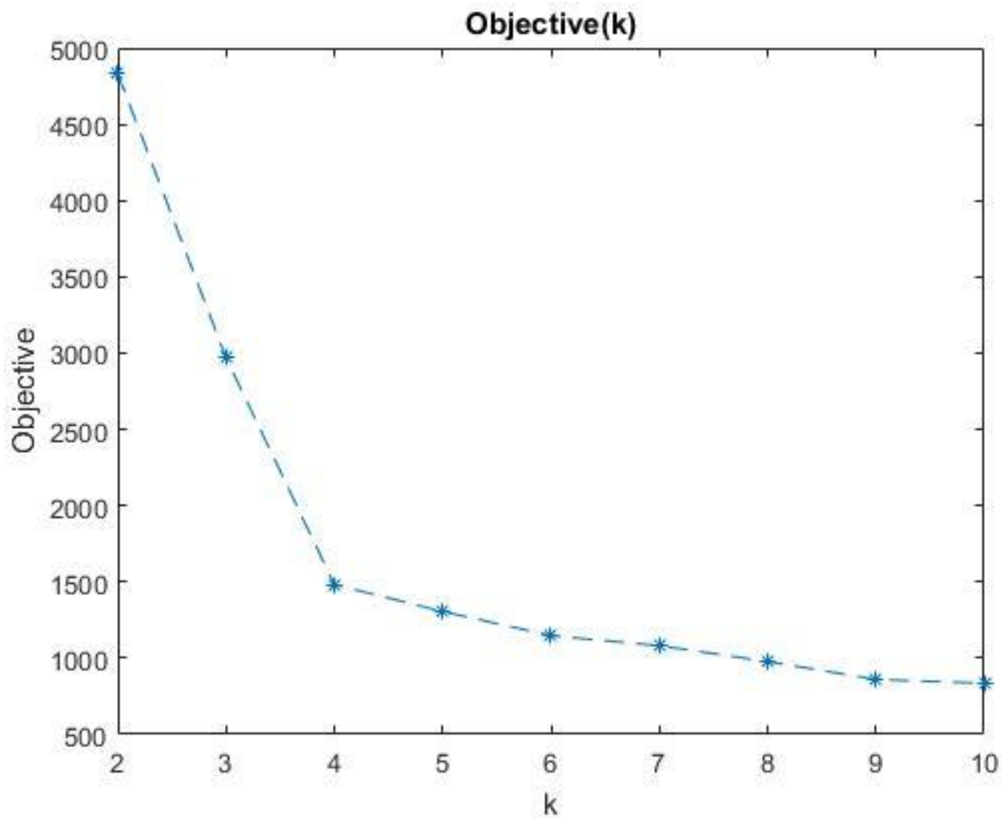
1.b.

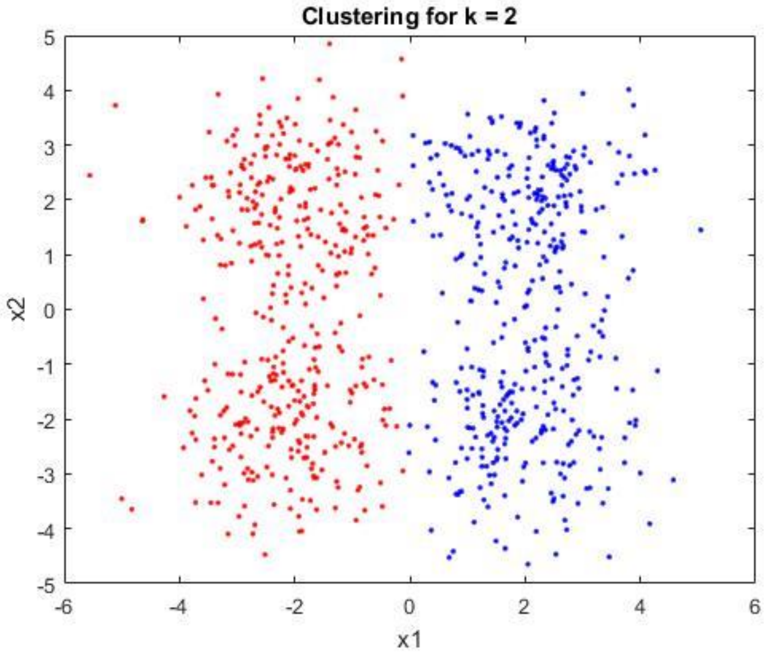


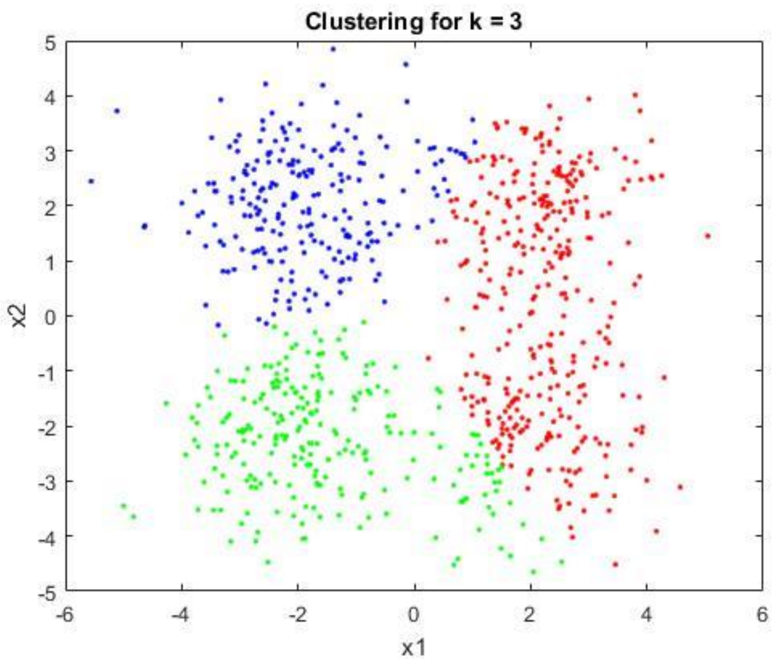
The objective decreases in k. This is logical if we look at the objective formula:

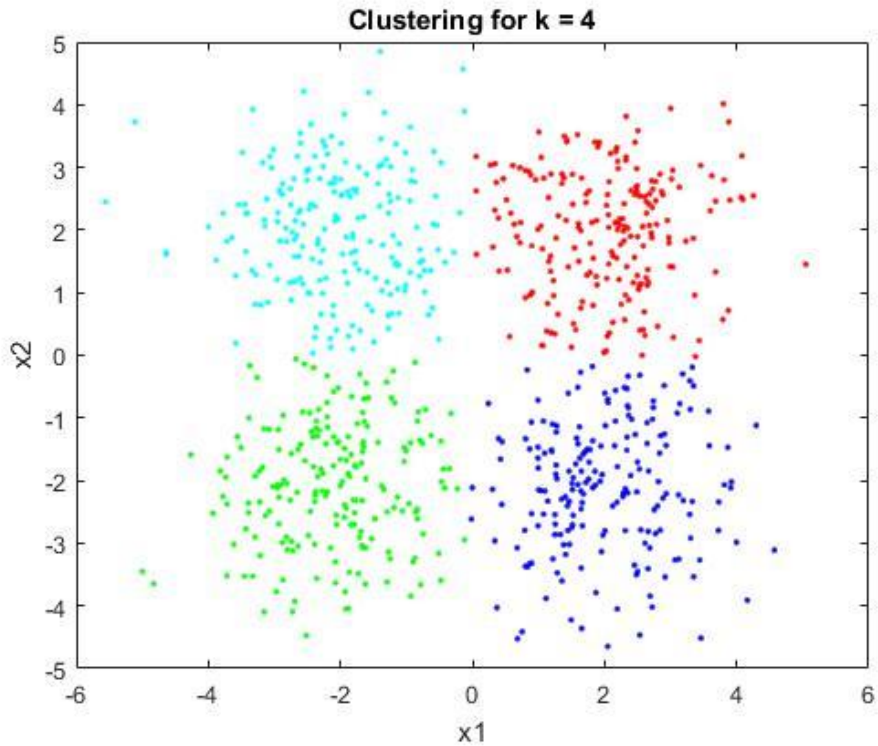
$$G_{k-m}(C_1, \dots, C_k) = \sum_{i=1}^{k} \sum_{x \in C_i} \|x - \frac{1}{|C_i|} \sum_{x \in C_i} x \|^2$$

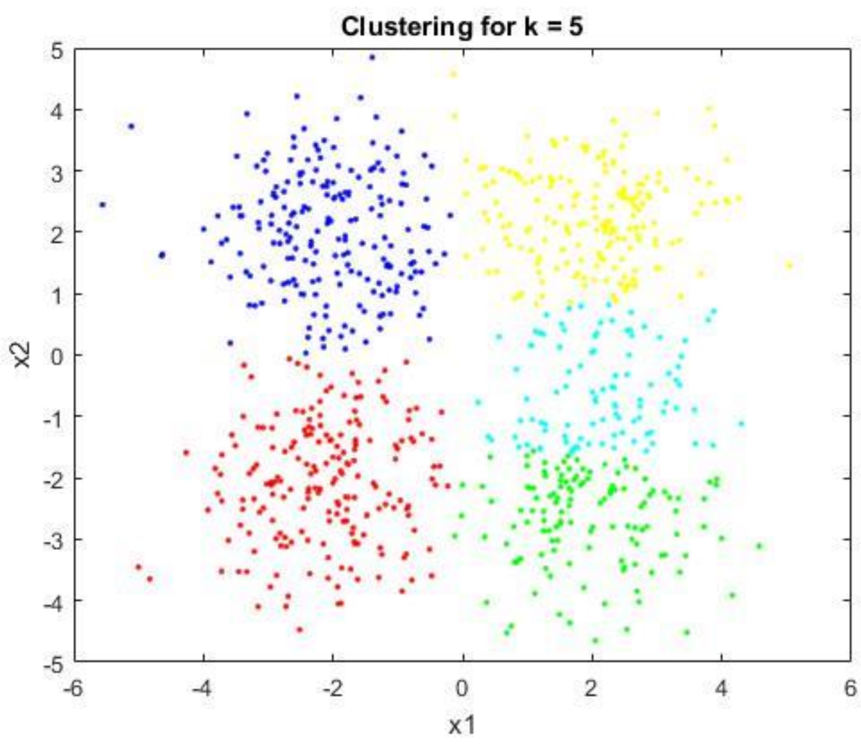
As k increases, each example has a center that is closer to it, so the summation of all distances will be smaller in the end.

1. c.


Clustering for k = 2
x2
x1
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6


Clustering for k = 3
x2
x1
5
4
3
2
1
0
-1
-2
-3
-4
-5
-6
-4
-2
0
2
4
6





By looking at the dataset, we can observe that the points generally divide into four groups. Therefore, k=4 is the most logical clustering (this can also be clearly seen in the figure for k=4). This is not the clustering with the smallest value of the k-means objective.

1.d.

By looking at the formula for the k-means objective, we can see that with high probability, the objective decreases in k regardless of the dataset. However, the most logical clustering is not always the one with the most number of clusters.

1.e.

Cluster	Most common label	Percentage of common label in Cluster (%)
1	7	91.0448
2	1	66.0606
3	6	71.1712
4	0	86
5	4	50
6	9	40.2878
7	8	54.8077
8	3	54.7445
9	0	91.8367
10	2	89.1892

1.f.

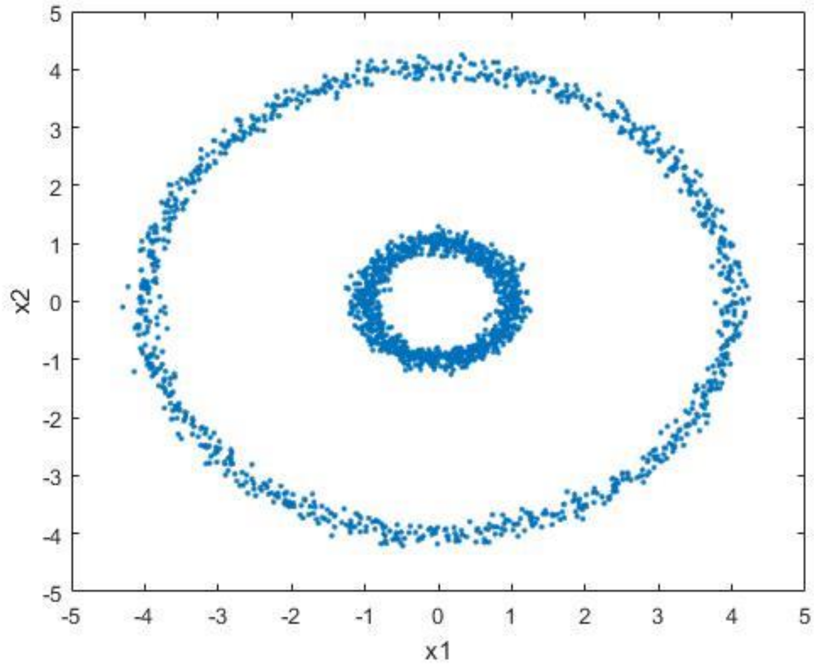
The classification error on the sample resulting from a classifier derived from the above clusters

Is: 0.357 this calculation was done by the following matlab calculation:

```
percentages = zeros(1, 10);
modes = zeros(1, 10);
Ypredict = zeros(size(C));
for i = 1:10
   labels = Y(C==i);
   modes(i) = mode(labels);
   percentages(i) = sum(labels==modes(i)) / size(labels,1) * 100.0;
   Ypredict = Ypredict + (C==i)*modes(i);
end
errs(j) = mean(Y ~= Ypredict);
```

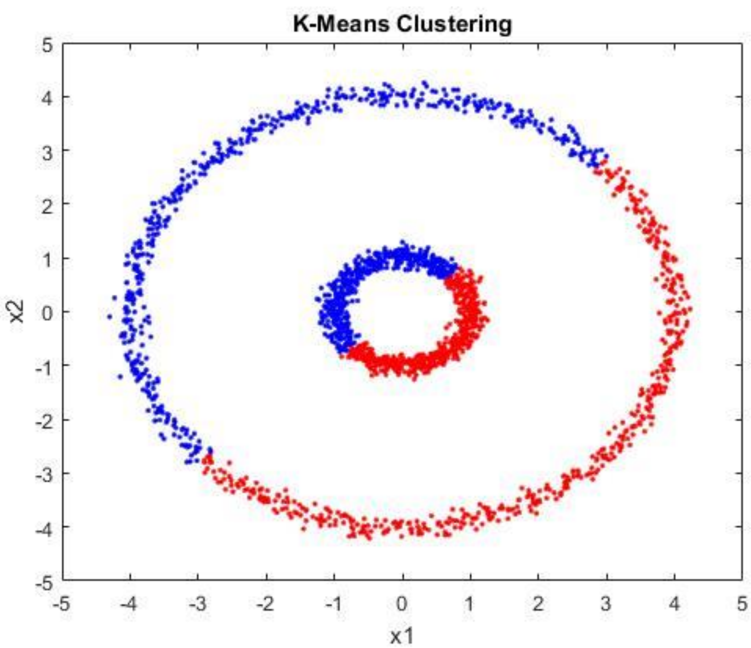
with C being the clusters array returned from the K-means, And Y being the original labels.

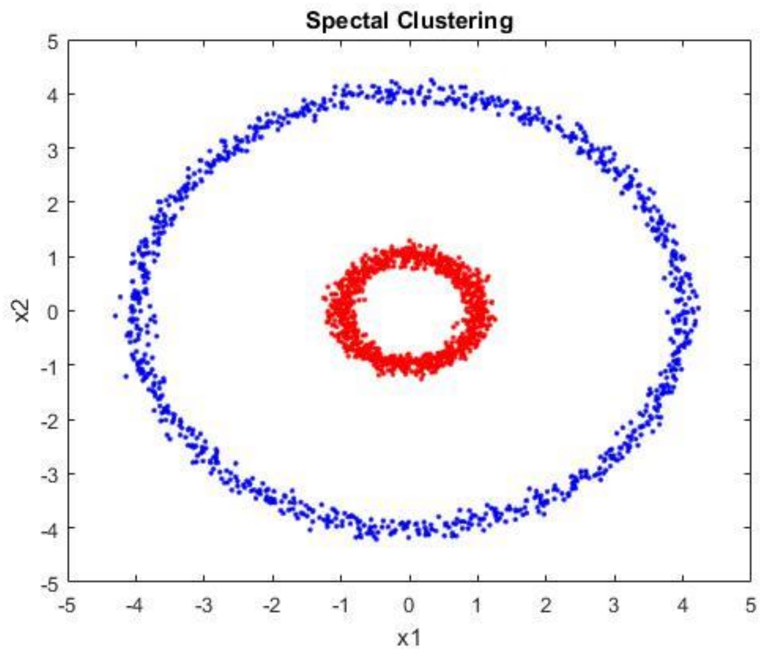
2.b.



k-means will find the clusters in which the points are closest to their center. In this case this will not output the desired outcome since the outer circle would never be identifies as one cluster.

2. c.





K-means Clustering tries to minimize the distances from each point to its cluster's center. Therefore, it "splits" the image in 2.
Spectral Clustering tries to minimize the variance of its clusters. Therefore, it identifies each circle as a cluster.

3.

③ $X=\mathbb{R}, Y=\mathbb{R}, y\sim N(3x, 5x^2)$

לפי ההגדרה, עלינו למצוא פונק' $h^*: X\to Y$ המביאה למינימום את $E_{(x,y)\sim D}[\ell(h(x),Y)]$. נחשב:

$$E_{(x,y)\sim D}(\ell(h(x),y)) = E_{(x,y)\sim D}((h(x)-y)^2) =$$

$$= E_x(E_y((h(x)-y)^2)) = E_x(E_y(h^2(x)-2h(x)\cdot y+y^2)) =$$

$$\left(\text{לינאריות התוחלת}\right) = E_x(E_y(h^2(x)) - E_y(2h(x)\cdot y) + E_y(y^2))$$

$$\left(\text{לינאריות התוחלת}\right) = E_x(h^2(x) - 2h(x)E_y(y) + E_y(y^2))$$

$$(*) = E_x(h^2(x) - 2h(x)\cdot 3x + 14x^2) = E_x(h^2(x) - 6x\cdot h(x) + 14x^2)$$

נמצא $h^*(x)$ המביאה את הביטוי $h^2(x)-6xh(x)+14x^2$ למינימום עבור כל $x\in\mathbb{R}$, ומכאן היא תביא את התוחלת לפי x של הביטוי למינימום:

$$g(h(x)) = h^2(x) - 6x\cdot h(x) + 14x^2$$

$$g'(h(x)) = 2h(x) - 6x = 0 \Rightarrow h^*(x) = 3x$$

$$g''(h^*(x)) = 2 > 0$$

לכן הפונקצ'ה $h^*(x)=3x$ היא Bayes-Optimal

(*) $Y\sim N(3x,5x^2)$, לכן $E_y(y)=3x$ ובנוסף,

$$E_y(y^2) = V(y) + E^2(y) = 5x^2 + 9x^2 = 14x^2$$

4.a. (4) א)

נוכיח שהבעיה קמורה:

נגדיר $B=\{x\in\mathbb{R}^d \mid Wx=Y\}$. נוכיח שהיא קב' קמורה:

יהי $u,v\in B$, $\alpha\in[0,1]$. נשים-לב כי $\alpha u+(1-\alpha)v\in\mathbb{R}^d$. בנוסף:

$$W(\alpha u+(1-\alpha)v)=\alpha Wu+(1-\alpha)Wv \underset{(u,v\in B)}{=} \alpha Y+(1-\alpha)Y=Y$$

לכן $\alpha u+(1-\alpha)v\in B$ ולכן B קמורה

נוכיח שהפונק' $f:B\to\mathbb{R}$ המוגדרת ע"י $f(w)=\|w\|_1$ קמורה:

$$f(\alpha u+(1-\alpha)v)=\|\alpha u+(1-\alpha)v\|_1 \underset{(\text{אי.ש. המשולש})}{\le} \|\alpha u\|_1+\|(1-\alpha)v\|_1$$

$$=\alpha\|u\|_1+(1-\alpha)\|v\|_1=\alpha f(u)+(1-\alpha)f(v)$$

לכן הבעיה קמורה ~~[illegible]~~ וניתן לפתור אותה באמצעות תוכנית לינארית מהצורה:

$$\max \langle u,x\rangle \quad \text{s.t. } Ax\ge r$$

נגדיר:

$$u_{2d\times 1}=(\overbrace{0,0,\dots,0}^{d};\overbrace{-1,-1,\dots,-1}^{d})$$

$$A_{(2k+2d)\times 2d}=\begin{pmatrix} W_{k\times d} & O_{k\times d}\\ -W_{k\times d} & O_{k\times d}\\ -I_{d\times d} & I_{d\times d}\\ I_{d\times d} & I_{d\times d}\end{pmatrix},\quad r_{2k+2d\times 1}=\begin{pmatrix} Y_{k\times 1}\\ -Y_{k\times 1}\\ O_{d\times 1}\\ O_{d\times 1}\end{pmatrix}$$

<u>הוכחת נכונות:</u> נסמן: $x=(x_1,x_2,\dots,x_d,\varepsilon_1,\varepsilon_2,\dots,\varepsilon_d)$.

נשים-לב כי התוכנית ממקסמת את $\langle u,x\rangle=\sum_{i=1}^d -\varepsilon_i$,

כלומר מביאה למינימום את $\sum_{i=1}^d \varepsilon_i$. ~~[illegible]~~ נגדיר $\bar{x}=(x_1,\dots,x_d)$ ונוכיח כי לכל $1\le i\le d$: $\varepsilon_i=|x_i|$ וגם $W\bar{x}=Y$:

לפי ה-constraints לכל פתרון מתקיים: $-W\bar{x}\ge -Y$ וגם $W\bar{x}\ge Y$

$\Rightarrow W\bar{x}\ge Y$ וגם $W\bar{x}\le Y$

$\Rightarrow W\bar{x}=Y$

בנוסף, מתקיים:

$$\forall 1\le i\le d: -x_i+\varepsilon_i\ge 0 \wedge x_i+\varepsilon_i\ge 0$$

$$\Rightarrow \forall i: \varepsilon_i\ge x_i \wedge \varepsilon_i\ge -x_i$$

$$\Rightarrow \varepsilon_i\ge \max\{x_i,-x_i\}=|x_i|$$

כיוון ואנחנו מביאים את ε_i למינימום, נקבל $\varepsilon_i=|x_i|$ ולכן התוכנית נכונה.

4.b.

(4ב) נוכיח שהפונקציה $f: B \to \mathbb{R}$ המוגדרת ע"י $f(x)=\|x\|_0$ אינה קמורה:

נגדיר: $d=2, k=1, W=(0,0), \gamma=0, u=(1,0), v=(1,1), \alpha=\frac{1}{2}$

נשים-לב כי $u,v \in \mathbb{R}^d$ וכן $Wu=Wv=\gamma$, לכן $u,v \in B$.

כמו-כן: $f(\alpha u+(1-\alpha)v) = \|\frac{1}{2}(1,0)+\frac{1}{2}(1,1)\|_0 = \|(1,\frac{1}{2})\|_0 = 2$

$\alpha f(u)+(1-\alpha)f(v) = \frac{1}{2}\|(1,0)\|_0 + \frac{1}{2}\|(1,1)\|_0 = \frac{1}{2}\cdot 1 + \frac{1}{2}\cdot 2 = \frac{3}{2}$

מתקיים $f(\alpha u+(1-\alpha)v) > \alpha f(u)+(1-\alpha)f(v)$

ולכן f אינה קמורה.

5.a.

נמצא את הבסיס שמוגדר ע"י מטריצת המעבר המקיימת: (5א)

$$A = \arg\min_{W\in\mathbb{R}^{d\times d},\, \det(A)\neq 0} \left(\sum_{i=1}^{m} \|Wx_i\|_0\right)$$

היתרון בשיטה זו הוא שאם נצטרך לשלוח את כל המדגם, המידע שניאלץ לשלוח יהיה מינימלי. החסרון הוא שאם נצטרך לשלוח x-ים מסויימים יותר פעמים מאחרים, זהו לא הפתרון האופטימלי (ניתן למצוא בסיס טוב יותר ע"י מתן משקל שונה ל-x-ים שונים)

5.b.

(5ב) ה-transmitter יחשב את מטריצת המעבר A ממרחב $\mathbb{R}^d$ למרחב של בסיס B ויחשב את Ax_i לכל $1\leq i\leq m$. כעת לכל $Ax_i(j)\neq 0$ הוא ישלח שלשה: $(i, j, x_i(j))$.

בהינתן המידע הנ"ל, ה-receiver ישחזר את המדגם במלואו בבסיס B. כעת הוא יבצע Compressed Sensing על מנת למצוא ייצוג דליל ל-S. כלומר, הוא יגריל מטריצת Grahm במימד kxd כאשר k הוא מימד תת-המרחב הנתון, ויכפיל אותה בכל אחד מאיברי המדגם. בהסתברות גבוהה (w.h.p) המטריצה שהוגרלה תהיה RIP ולכן המידע יעבור לייצוג דליל $(m\times k)$ אשר נוכל לשחזר ממנו את S כולה.

נשים-לב שחוץ משגיאות בשליחה, השגיאה יכולה להיות רק בהגרלת מטריצת ה-RIP, וזו אותה השגיאה שעלולה לקרות באלגו' אופטימלי שבו ה-receiver יודע את S ומנסה להעבירו לייצוג דליל.

6.a.

(6) a) נדרוש ש- $f_\lambda(x)$ תקיים את התכונות של פונק' צפיפות:

$$\int_{-\infty}^{\infty} f_\lambda(x)\,dx = 1$$

$$\Rightarrow \int_{0}^{\infty} C_\lambda e^{-\lambda x}\,dx = 1$$

$$C_\lambda \int_{0}^{\infty} e^{-\lambda x}\,dx = 1$$

$$C_\lambda \left(\frac{e^{-\lambda x}}{-\lambda} \Big|_0^{\infty} \right) = 1$$

$$C_\lambda \left(0 + \frac{1}{\lambda} \right) = 1$$

$$\boxed{C_\lambda = \lambda}$$

(6) b) $f_\lambda(x) = \lambda e^{-\lambda x}$

$$L(S;\lambda) = \sum_{i=1}^{m} \log f_\lambda(x_i) = \sum_{i=1}^{m} \log\left(\lambda e^{-\lambda x_i}\right) = \sum_{i=1}^{m} \left(\log\lambda + \log e^{-\lambda x_i}\right)$$

$$= m\log\lambda - \lambda \sum_{i=1}^{m} x_i$$

נגזור כדי למצוא מקסימום:

$$L'(S;\lambda) = \frac{m}{\lambda} - \sum_{i=1}^{m} x_i = 0$$

$$\Rightarrow \lambda = \frac{m}{\sum_{i=1}^{m} x_i}$$

נוכיח שזוהי נק' מקסימום:

$$L''(S;\lambda) = -\frac{m}{\lambda^2} < 0$$

לכן ה-λ שמצאנו ממקסמת את $L(S;\lambda)$ ולכן:

$$\boxed{\hat{\lambda} = \frac{m}{\sum_{i=1}^{m} x_i}}$$