מבוא לאנליזה נומרית - ש.ב. 1
אמיר ארבל, גלעד צודיקוביץ

שאלה 1

א.

סימן	חזקה	שבר
1	10001000	00100010101100110011010

ב. מספר המספרים השלמים הניתנים לייצוג בעזרת IEEE754:
מתוך כל האפשרויות עבור המנטיסה, בחצי מהספרות ה1 הכי ימני יהיה הביט האחרון, עבור כל סיפרה כזו נקבל מספר שלם אם נכפיל אותו ב2 בחזקת X כך X מספר שלם בין 23 ל127 כולל.
נחזור על הפעולה עבור כל מספר שה1 הכי ימני בו הוא הביט האחד לפני האחרון (ברבע מהאפשרויות) כעת ניתן לקבל מספר שלם אם נכפיל ב2 בחזקת X כך שX מספר שלם בין 22 ל127 כולל.
נחזור על הפעולה עד שה1 הכי ימני יהיה הביט הראשון (מקרה בודד) אותו ניתן להכיל ב2 בחזקת X כך שX מספר שלם בין 2 ל127 כולל.
נכפול ב2 עבור כל אפשרות של סיבית הסימן.

$$2 \cdot \sum_{i=1}^{23} 2^{23-i} \cdot (127 - 23 + (i+1))$$

ג. 6 ספרות בהצגה דסימלית. דבר זה נובע מהפרשים בין חזקות של 2 שניתן ליצג - הרי ההפרש הכי גדול הוא $2^{-23} = \frac{1}{8388604}$. כאשר עוברים את החזקה ה-23, נוצר פער בין מספרים שניתן ליצג, לכן הדיוק יורד ככל שהמספר עולה.

ד.

i) $3.4028 \cdot 10^{38}$

ii) $1.18 \cdot 10^{-38}$

ה) 2^{23}

שאלה 2

א.

```
function [est_pi, error, d] = calcPi(n)
    pi_estimate = 2^n;
    last = 0;
    while (n > 1)
        last = sqrt(vpa(2 + last));
        n = n - 1;
    end
    last = sqrt(2-last);
    est_pi = pi_estimate * last;
    error = vpa(pi) - vpa(est_pi);
    d = ceil(log10(error));
return
```

ב.

n	d
4	2
8	4
12	7
16	9
20	11
24	14

ג. הסטודנט לא צדק, כי מתקיים $lim_{n \to \infty} \sqrt{2+\sqrt{2+\sqrt{2+...}}} = 2$. כלומר לח גדול מספיק, הערך של $lim_{n \to \infty} \sqrt{2+\sqrt{2+\sqrt{2+...}}} - 2$ הוא קטן מאוד ולא מחושב נכון - המחשב יגיע לתוצאה 0 שתוביל לשגיאה בחישוב.

ד. נוסיף דיוק:

```
function [est_pi1, error1, d1] = calcPi1(n)
  digits(1000)
  format LONG
  [est_pi1, error1,d1] = calcPi(n);

return
```

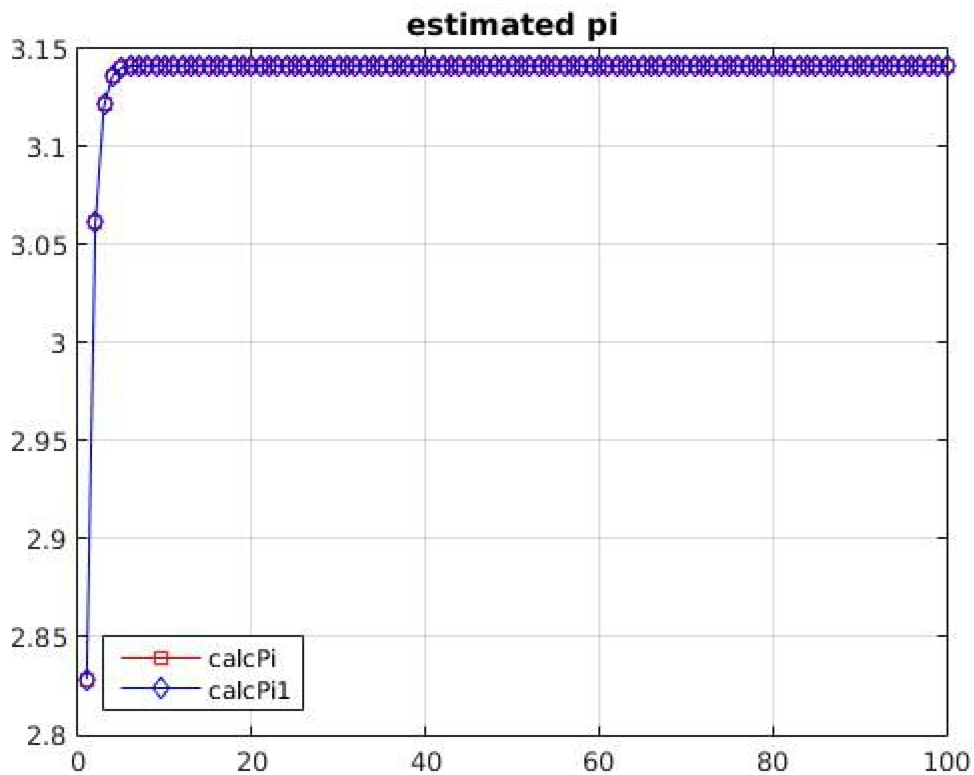
ה.

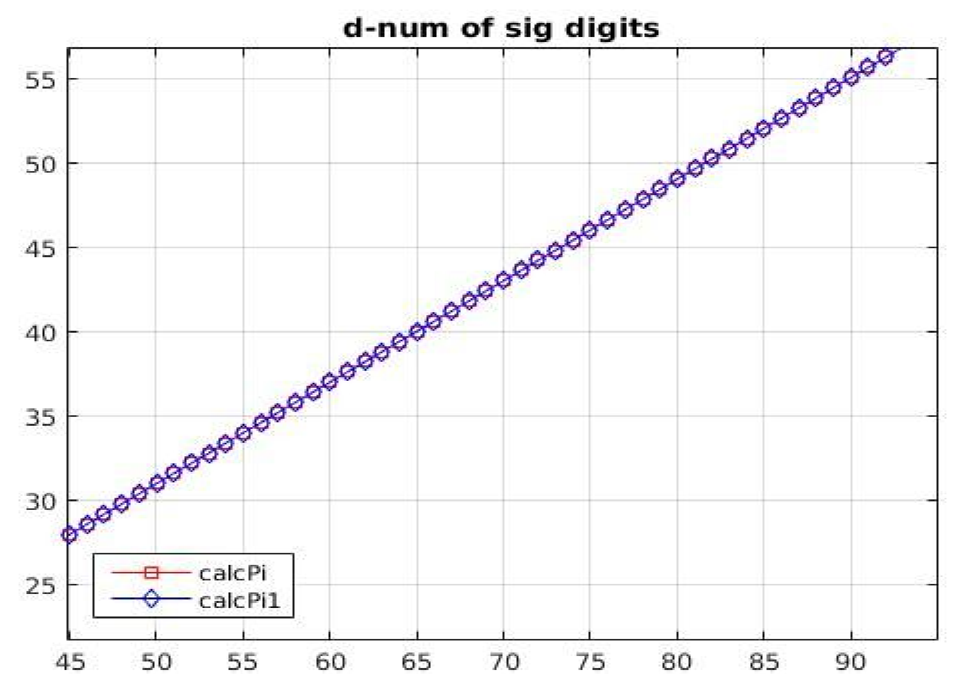
n	d
4	2
8	4
12	7
16	9
20	11
24	14

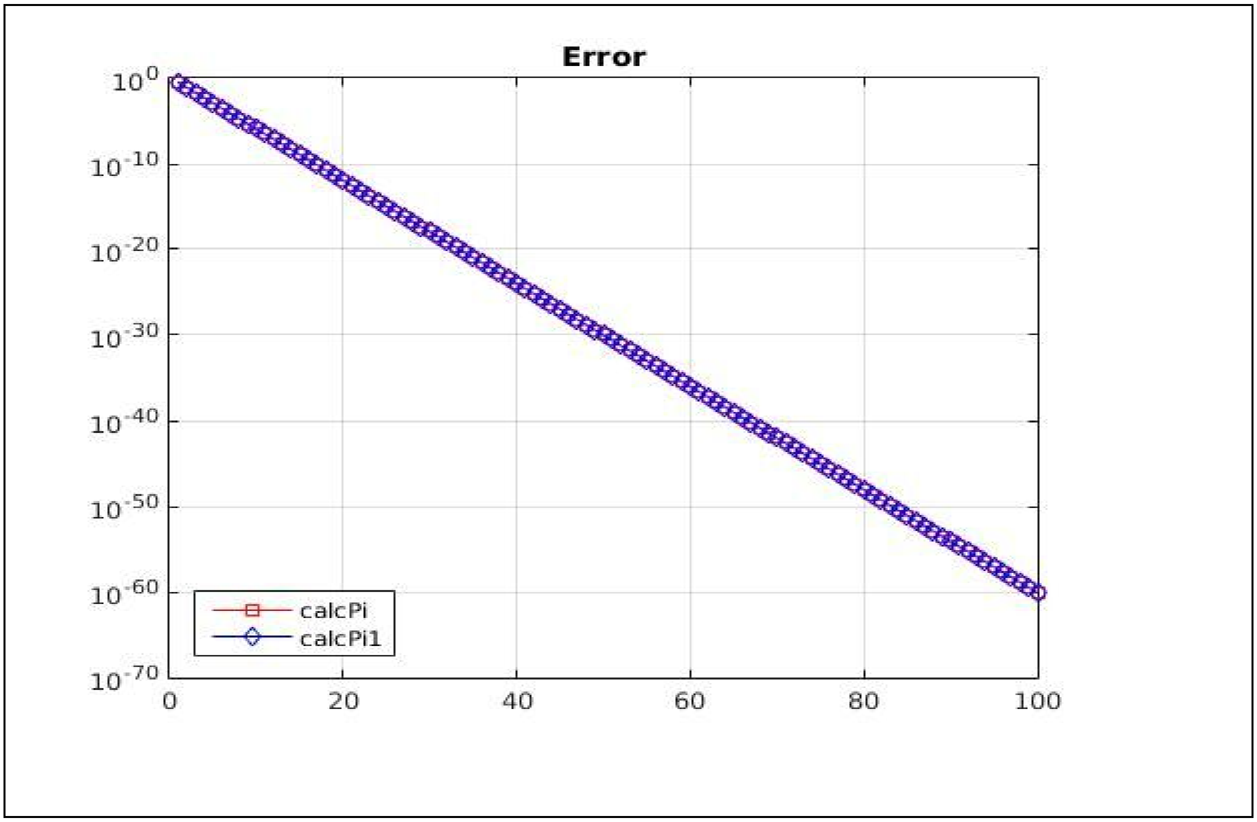
ו. הקוד להצגת הגרפים:

```
t=100;
y1=zeros(1,t);
y2=zeros(1,t);
for n=1:t
    [est_pi,err,d]=calcPi(n);
    e1(1,n)=err;
    d1(1,n)=d;
    p1(1,n)=est_pi;
    [est_pi,err,d]=calcPi1(n);
    e2(1,n)=err;
    d2(1,n)=d;
    p2(1,n)=est_pi;
end

x=1:t;
figure(1)
semilogy(x,e1,'rs-');
hold on

semilogy(x,e2,'bd-');
grid on
hold off
figure(2)
plot(x,d1,'rs-');
hold on
plot(x,d2,'bd-');
grid on
hold off

figure(3)
plot(x,p1,'rs-');
hold on
```

```
plot(x,p2,'bd-');
hold off
```







3. נסמן:

$$x = y_1.y_2...y_n...y_k \cdot \beta^e$$
$$\tilde{x} = x_1.x_2...x_n \cdot \beta^e$$

נשים לב כי:

$$\beta - Base, \; n \le k, \; m \le e \le M$$

אם כן, לפי משפט שלמדנו, מספר הספרות המשמעותיות (בחיתוך) של $\tilde{x}$ הוא d אם מתקיים ש-d הוא המספר השלם האי-שלילי הקטן ביותר המקיים:

$$\delta\tilde{x} = |x - \tilde{x}| / |x| \le \beta^{1-d}$$

נרצה להראות: d=n:

$$\delta\tilde{x} = |x - \tilde{x}| / |x| = \frac{|y_1.y_2...y_n...y_k \cdot \beta^e - x_1.x_2...x_n \cdot \beta^e|}{|y_1.y_2...y_n...y_k \cdot \beta^e|}$$

מכיוון ש-$\tilde{x}$ הוא קירוב בחיתוך של x ל-n ספרות, מתקיים $1 \le \forall i \le n$ $x_i = y_i$. אם כן:

$$\frac{|y_1.y_2...y_n...y_k \cdot \beta^e - x_1.x_2...x_n \cdot \beta^e|}{|y_1.y_2...y_n...y_k \cdot \beta^e|} = \frac{|0.0...0y_{n+1}...y_k \cdot \beta^e|}{|y_1.y_2...y_n...y_k \cdot \beta^e|} = \frac{|y_{n+1}.y_{n+2}...y_k \cdot \beta^{e-n}|}{|y_1.y_2...y_n...y_k \cdot \beta^e|} \le \beta^{1-n}$$

מכיוון ש-d הוא המספר השלם האי-שלילי הגדול ביותר המקיים זאת, ניתן להניח ש-$d \ge n$. נניח בשלילה כי $d > n$: לדוגמה, $d = n + 1$, ונסתכל על המקרה הבא:

$$\beta = 10, \; n = 4, \; x = 1.4321567 \cdot 10^6, \; \tilde{x} = 1.432 \cdot 10^6$$

אם כן:

$$\delta\tilde{x} = \frac{\left|1.4321567\cdot10^6 - 1.432\cdot10^6\right|}{\left|1.4321567\cdot10^6\right|} = \frac{\left|1.567\cdot10^2\right|}{\left|1.4321567\cdot10^6\right|} \geq 1\cdot10^{-4} = 10^{1-(n+1)}$$

בסתירה להנחה.

4.
א.

```
function sum = anss(array)
            sum = 0;
            for ii = 1:numel(array)
               sum = sum + array(ii);
            end
```

נוכיח כי פונקציה זאת תעבוד לדיוק יחסי של 10^{-7} במקרה שנקבל מערך בגודל $2^n, 2 \leq n \leq 20$ אשר איבריו הם $1 \leq c \leq 1000$ ויצוגם כSingle הוא $\tilde{c}$:
נזכור כי לפי שאלה קודמת, Single מחזיק ב6 ספרות מדויקות לפחות. מכיוון שהמספר חסום על ידי 1000.
כלומר הדיוק לכל מספר הוא לכל הפחות 10^{-2} . נסתכל על המקרה שבו x מקסימלי קטן ככל האפשר עם שגיאה כדי להגדיל את השגיאה היחסית. מקרה קרוב הוא:
x=1+10^-8 $\tilde{x} = 1$
באופן טריוואלי, השגיאה היחסית תהיה קטנה מ- 10^{-7} , לכן נסתכל על הקצה השני (במקרה הגרוע).
כאן החזקה הבסיסית היא $2^9 = 512$, לכן הדיוק הוא עד 10^{-4} למעשה. אם כן
$x = 999 + 9\cdot10^{-5}$ $\tilde{x} = 1000$

אם כן השגיאה היחסית:

$$\delta\tilde{x} = \left|x - \tilde{x}\right| / |x| \leq \left|2^n\cdot10^{-4}\right| / \left|2^n\cdot999.00009\right| = 10^{-4}/(9.9900009\cdot10^3) \leq 10^{-7}$$

ב. חסם מקורב על השגיאה בסכום של 10^6 ערכים בין 1 ל-2:
לערכים בין 1 ל-2 במשתני single יש שגיאה של לכל היותר $\frac{1}{2^{23}} = \frac{1}{8688608}$. אם כן למיליון ערכים כאלו תהיה שגיאה של לכל היותר $0.12 \sim \frac{10^6}{2^{23}} = \frac{1000000}{8688608}$, והשגיאה היחסית תהיה לכל היותר $\frac{1}{8688608}$.
ג.

```
function sum = anss(array)
   sum = double(0);
   for ii = 1:numel(array)
      sum = sum + array(ii) ;
   end
   array= single(array);
   array= double(array);
   for ii = 1:numel(array)
      sum = sum - array(ii) ;
```

```
end
```

5. $f(x) = x^2,\ \tilde{x} = x + \Delta x,\ x,\ \tilde{x} > 0$.

אם כן:

$$f(\tilde{x}) = f(x + \Delta x) = f(x \pm 1) = (x\ +\ 1)^2 = x^2 \pm (2x + 1)\,.$$

ידוע כי בכפל השגיאה: $\delta(\tilde{x}\ \cdot \tilde{y}) = \delta\tilde{x} + \delta\tilde{y}$ מכאן השגיאה בכפל היא:

$$\delta(\tilde{x}\ \cdot \tilde{x}) = \delta\tilde{x} + \delta\tilde{x} = 2\frac{\Delta x}{x}$$

נשים לב של $\Delta x\ \geq\ 1$ החסם לא מתקיים, כאשר בעצם הנחנו ש $\tilde{x} \cdot \tilde{x}$ הוא מספר קטן ביותר (אשר מתכנס 0), אך לא ניתן להניח זאת כאשר בכפל הוא לא קטן - כלומר כאשר $\Delta x\ \geq\ 1$.

והשגיאה לפי גרדיאנט:

$$\delta(\widetilde{x^2}) = \frac{(x^2)'\,\Delta x}{x^2} = 2\frac{\Delta x}{x}$$