

שיטות נומריות 88-376

מרצים של הקורס : דר' אגרנוביץ' אלכסנדרה, מר אלתר עידן

שעות קבלה : בתיאום

מייל : alexandra.agranovich@gmail.com
alterid@cs.biu.ac.il

מפרט טכני

- אתר הקורס: ב-Moodle (lemida.biu.ac.il)

- תרגילי בית: 6 תרגילים

- הרכב הציון סופי בקורס:

מבחן	בוחן
85%	15%

- תאריך קיום הבוחן : 3.08

למה יש צורך בשיטות נומריות?

- פתרון אנליטי לא תמיד אפשר למצוא.

דוגמא : מציאת אינטגרל $\int_0^2 e^{-x^2} dx$

- מציאת פתרון אנליטי דורש זמן רב ולא נחוץ לפתרון הבעיה

- עיבוד אנליטי של תוצאות המדידה

- דוגמא : מציאת גובה הטיסה של הטיל כפונקציה של הזמן.

רקע

- שיטה נומרית טובה מקיימת את שלושת התכונות הבאות :

1. דיוק – הפתרון צריך להיות ככל שקרוב יותר לפתרון המדויק

2. איכות – האלגוריתם נדרש לספק את פתרונות לבעיות רבות

ולהיות מסוגל לאמוד את שיעור השגיאה

3. מהירות – סיבוכיות האלגוריתם

ניתוח השגיאות

Error Analysis

אוניברסיטת בר-אילן, תשע"ו. מאת : דר' אלכסנדרה אגרנוביץ'

מקורות השגיאה

- **שגיאות קיצוץ / עיגול כתוצאה מחישובים (Round-off errors) :**
עיגול מספרים אי-רציונליים ידנית או במחשב, המרה של שבר פשוט למספר עשרוני במחשב
- **שגיאת קירוב/שגיאת אלגוריתם (Truncation Errors) :** לקיחת מספר סופי של איברים בטור אינסופי, קירוב של פונקציה לא ליניארית ע"י פונקציה ליניארית, חישוב נגזרת בעזרת מנת הפרשים
- **שגיאה מתפשטת (Propagation error) :** שגיאה בפלט כתוצאה של שגיאה בקלט

מדידת שגיאות. הגדרות.

יהי x^* ערך מקורב של ערך אמיתי x . אזי

- שגיאה מוחלטת : $\Delta = |x - x^*|$

- שגיאה יחסית: $\delta = \left| \frac{x - x^*}{x} \right|$ עבור $x \neq 0$. שגיאה יחסית גם

נמדדת באחוזים.

- לרוב מחפשים חסם לשגיאה

דוגמא

נניח שקירבנו את $x = \sqrt{2}$ בעזרת 1.41. מהי השגיאה המוחלטת ומהי השגיאה היחסית? עבדו עם מספרים בעלי 5 ספרות משמעותיות.

$$\sqrt{2} \approx 1.414213562373095$$

פתרון:

$$\Delta x = \left| \sqrt{2} - 1.41 \right| \leq 0.0042136$$

$$\delta x = \frac{\Delta x}{x} = \left| \frac{\sqrt{2} - 1.41}{\sqrt{2}} \right| \leq \frac{0.0042136}{1.4142} \approx 0.0029795 \approx 0.3\%$$

צורה כללית

- עבור נקודות וקטוריות :

$$\Delta = \left\| \vec{x} - \vec{x}_0 \right\|$$

$$\delta = \frac{\left\| \vec{x} - \vec{x}_0 \right\|}{\left\| \vec{x} \right\|}$$

- דוגמא שלנורמה 2 . עבור $\mathbf{x} = (x_1, \dots, x_n)^T$: $\left\| \mathbf{x} \right\|_2 = \sqrt{\sum_{i=1}^n x_i^2}$

שגיאות עיגול/קיצוץ

- ייצוג מספרים

- פורמט נקודה קבועה (fixed point) - ישנו מספר קבוע של ספרות אחרי

נקודה

- פורמט נקודה צפה (floating point) - עובדים עם מספר קבוע של

ספרות (הנקודה יכולה "לזוז").

- ייצוג מספרים בפורמט של נקודה צפה בצורה כללית

$$x = \sigma \cdot (m)_B \times B^e$$

כאשר $\sigma = \pm 1$

e - מספר שלם ($n \leq e \leq N$)

צורה תקינה של מנמיסה בעלת p ספרות ב-floating point היא

$$d_0.d_1\dots d_{p-1}$$

כאשר $d_0 \neq 0, 0 \leq d_i < B$

ייצוג מספרים במחשב

תקן Float (Single Precision) : 32 סיביות למספר

תקן Double (Double Precision) : 64 סיביות למספר

1	2-24	25-32
	2-53	54-64

שדה

שדה מנטיסה

שדה מעריך

סימן

עבור תקן Float :

- החזקה הנמוכה ביותר היא 126 – והייצוג שלה הוא: 0000 0001
- החזקה הגבוהה ביותר היא 127 והיצוג שלה הוא: 1111 1110
- החזקה המחשבת 1111 1111 מייצגת את ∞
- המספר 0 מיוצג ע"י מנטיסה וחזקה השוות כולן לאפסים

שגיאת עיגול/ קיצוץ (Floating point)

נסמן ייצוג בצורת נקודה צפה של מספר x במחשב ע"י $fl(x)$

• שגיאה מוחלטת בייצוג מספר במחשב

$$|x - fl(x)| \leq \begin{cases} \frac{1}{2} \cdot B^{e-p+1}, & \text{for rounding} \\ B^{e-p+1}, & \text{for chopping} \end{cases}$$

כאשר p - אורך המנטיסה

הוכחה במקרה של עיגול (Floating point)

נזכר כי $fl(x) = \pm d_0.d_1 \dots d_{p-1} \cdot B^e$, לכן

$$\Delta x = |x - fl(x)| = |x - d_0.d_1 d_2 \dots d_{p-1} \cdot B^e| \leq \underbrace{0.00 \dots 0}_p B' \cdot B^e$$

$B' = \frac{B}{2}$ (עיגול B למספר הקרוב ביותר)



$$\Delta x \leq \frac{B}{2} \cdot B^{-p} \cdot B^e = \frac{1}{2} B^{1-p+e}$$

• שגיאה יחסית בייצוג מספר במחשב

$$\left| \frac{x - fl(x)}{x} \right| \leq \begin{cases} \frac{1}{2} \cdot B^{1-p}, & \text{for rounding} \\ B^{1-p}, & \text{for chopping} \end{cases}$$

הוכחה במקרה של עיגול (Floating point)

$$\delta x = \left| \frac{\Delta x}{x} \right| \leq \frac{\frac{1}{2} B^{1-p+e}}{d_0 \cdot d_1 d_2 \dots d_{p-1} \cdot B^e} \underset{(d_0 \neq 0)}{\leq} \frac{1}{2} \cdot \frac{B^{1-p+e}}{B^e} = \frac{1}{2} \cdot B^{1-p}$$

ϵ - מכונה

- בכל מחשב אפשר להגדיר דיוק של הצגה של מספר. יש כמה דרכים להגדיר אותו. הדרך הכי פופולארית היא ע"י ϵ - מכונה. הוא מוגדר כהפרש בין 1 למספר הבא הגדול מ-1 שאפשר להציג במחשב;

- בפורמט single precision (float) במחשב, למנטיסה יש 23 סיבים, לכן המספר הבא שאפשר לשמור הוא $1 + 2^{-23}$ ולכן ϵ - מכונה הוא $2^{-23} = 1.19 \times 10^{-7}$

- בפורמט double : ϵ - מכונה הוא $2^{-52} = 2.22 \times 10^{-16}$

שגיאת קירוב

- באנליזה נומרית, שימוש בטור טיילור הוא אחת הדרכים לחשב ערך מקורב של פונקציה נתונה

טור טיילור

תהי $f(x)$ פונקציה הגזירה $n+1$ פעמים בסביבה U של הנקודה x_0 .
אזי לכל $x \in U$ קיימת נקודה c הנמצאת בין x_0 ל- x כך שמתקיים:

$$f(x) = P_n(x) + R_n(x) = \sum_{k=0}^n \frac{f^{(k)}(x_0)}{k!} (x - x_0)^k + R_n(x)$$

כאשר

$$R_n(x) = \frac{f^{(n+1)}(c)}{(n+1)!} (x - x_0)^{n+1}$$

שגיאת מתפשטת (פעולות חשבון)

אם חישוב נעשה עם מספרים שהתקבלו עם שגיאה, התוצאה תושפע מהשגיאות האלה. נסתכל קודם על השפעתם על תוצאות של פעולות אריתמטיות. נסמן x, y - מספרים אמיתיים ו- x^*, y^* - מספרים שהשתתפו בחישובים. רוצים להעריך שגיאה מתפשטת

$$E = (x^* \omega y^*) - (x \omega y)$$

כאשר אומגה מסמלת 4 פעולות חשבון.

$$\text{נניח כי } x^* = x + \varepsilon_1, y^* = y + \varepsilon_2$$

$$\text{חיבור : } x^* + y^* = (x + y) + (\varepsilon_1 + \varepsilon_2) \Rightarrow E_+ = \varepsilon_1 + \varepsilon_2$$

$$\text{כפל : } x^* y^* = xy + x\varepsilon_2 + y\varepsilon_1 + \varepsilon_1\varepsilon_2 \Rightarrow E_* = x^* y^* - xy = x\varepsilon_2 + y\varepsilon_1 + \varepsilon_1\varepsilon_2$$

אזי שגיאה יחסית היא

$$\text{Rel}(xy) = \frac{E_*}{xy} = \frac{x\varepsilon_2 + y\varepsilon_1 + \varepsilon_1\varepsilon_2}{xy} = \frac{\varepsilon_2}{y} + \frac{\varepsilon_1}{x} + \frac{\varepsilon_1\varepsilon_2}{xy} = \text{Rel}(y) + \text{Rel}(x) + \text{Rel}(y) \cdot \text{Rel}(x)$$

בהנחה ש- $|\text{Rel}(y)|, |\text{Rel}(x)| \ll 1$, נקבל $\text{Rel}(xy) \approx \text{Rel}(y) + \text{Rel}(x)$. תוצרה טובה מאוד ו

שגיאת מתפשטת (פונקציות)

- כתוצאה משגיאה בקלט חישובנו $f(x^*)$ במקום $f(x)$. בהנחה ש- f גזירה ברציפות סביב x^* , אפשר לרשום

$$f(x) - f(x^*) \approx f'(x^*)(x - x^*)$$

- קיבלנו $E(f(x)) \approx f'(x^*)(x - x^*)$ וגם

$$\text{Rel}(f(x)) \approx \frac{f'(x^*)(x - x^*)}{f(x)} \approx \frac{f'(x)(x - x^*)}{f(x)} = \frac{f'(x)}{f(x)} x \text{Rel}(x)$$

- עבור פונקציה של n משתנים

$$E \approx \sum_{i=1}^n \frac{\partial f}{\partial x_i}(x^*)(x_i - x_i^*) \Rightarrow \Delta f = |E| \leq \sum_{i=1}^n \left| \frac{\partial f}{\partial x_i}(x^*) \right| \cdot |x_i - x_i^*| = \sum_{i=1}^n \left| \frac{\partial f}{\partial x_i}(x^*) \right| \cdot \Delta x_i$$

שגיאת מתפשטת

$$\Delta f \leq \sum_{i=1}^n \left| \frac{\partial f}{\partial x_i} \right| \cdot \Delta x_i$$

תרגיל

מהן השגיאות של סכום, הפרש, מכפלה וחילוק של 2 מספרים?

פתרון

$$\Delta(x + y) \leq \left| \frac{\partial(x + y)}{\partial x} \right| \cdot \Delta x + \left| \frac{\partial(x + y)}{\partial y} \right| \cdot \Delta y = |1| \cdot \Delta x + |1| \cdot \Delta y = \Delta x + \Delta y$$

$$\Delta(x - y) \leq \left| \frac{\partial(x - y)}{\partial x} \right| \cdot \Delta x + \left| \frac{\partial(x - y)}{\partial y} \right| \cdot \Delta y = |1| \cdot \Delta x + |-1| \cdot \Delta y = \Delta x + \Delta y$$

$$\Delta(x \cdot y) \leq \left| \frac{\partial(x \cdot y)}{\partial x} \right| \cdot \Delta x + \left| \frac{\partial(x \cdot y)}{\partial y} \right| \cdot \Delta y = |y| \cdot \Delta x + |x| \cdot \Delta y$$

$$\Delta\left(\frac{x}{y}\right) \leq \left| \frac{\partial\left(\frac{x}{y}\right)}{\partial x} \right| \cdot \Delta x + \left| \frac{\partial\left(\frac{x}{y}\right)}{\partial y} \right| \cdot \Delta y = \left| \frac{1}{y} \right| \cdot \Delta x + \left| -\frac{x}{y^2} \right| \cdot \Delta y = \frac{1}{|y|} \cdot \Delta x + \frac{|x|}{y^2} \cdot \Delta y$$

$$\Delta f \leq \sum_{i=1}^n \left| \frac{\partial f}{\partial x_i} \right| \cdot \Delta x_i$$

$$\Delta(x \cdot y) \leq |y| \cdot \Delta x + |x| \cdot \Delta y \quad \text{בפרק הקודם :}$$

$$\Delta\left(\frac{x}{y}\right) \leq \frac{1}{|y|} \cdot \Delta x + \frac{|x|}{y^2} \cdot \Delta y$$

מכאן,

$$\delta(xy) = \frac{\Delta(xy)}{|xy|} \leq \frac{|y|\Delta x + |x|\Delta y}{|xy|} = \frac{\Delta x}{|x|} + \frac{\Delta y}{|y|} = \delta x + \delta y$$

$$\delta\left(\frac{x}{y}\right) = \frac{\Delta\left(\frac{x}{y}\right)}{\left|\frac{x}{y}\right|} \leq \frac{\frac{1}{|y|} \cdot \Delta x + \frac{|x|}{y^2} \cdot \Delta y}{\left|\frac{x}{y}\right|} = \frac{\Delta x}{|x|} + \frac{\Delta y}{|y|} = \delta x + \delta y$$

התבטלות

(loss of significance/ catastrophic cancellation)

דוגמא 1 : נניח שהמחשב שלנו עובד עם מספרים בעלי 6 ספרות בשיטת עיגול. צריך לחשב $f(500)$ עבור $f(x) = x(\sqrt{x+1} - \sqrt{x})$

נחשב ישירות ונקבל

$$f(500) = 500(\sqrt{501} - \sqrt{500}) = 500(22.3830 - 22.3607) = 500(0.0223) = 11.1500$$

נכתוב את הפונקציה בצורה אחרת

$$g(x) = \frac{x(\sqrt{x+1} - \sqrt{x})(\sqrt{x+1} + \sqrt{x})}{(\sqrt{x+1} + \sqrt{x})} = \frac{x}{(\sqrt{x+1} + \sqrt{x})}$$

$$g(500) = \frac{500}{(\sqrt{501} + \sqrt{500})} = \frac{500}{(22.3830 + 22.3670)} = 11.1748$$

התשובה המדויקת היא 11.1747553...

מכאן, אפשר ללמוד איך עדיף להציג נוסחת השורשים בפתרון של משוואה ריבועית

התבטלות

(loss of significance/ catastrophic cancellation)

$$\text{דוגמא 2 : } f(x) = \frac{1 - \cos x}{x^2}, \quad f(0.01)$$

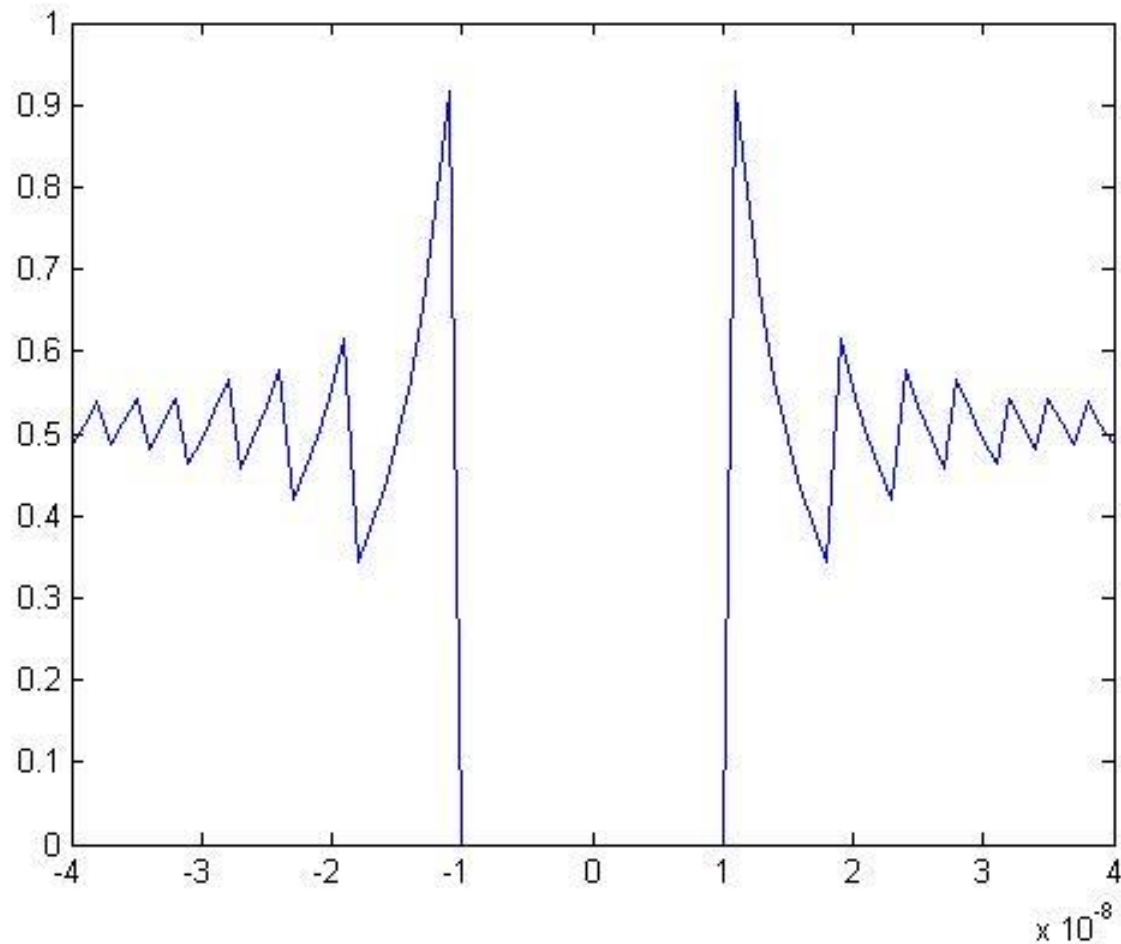
בעיה : בחישוב ישיר אפשר לקבל שגיאה גדולה כי 0.01 קרוב מאוד ל-0 ולכן קוסינוס שלו קרוב ל-1

הפתרון : לפתח את הקוסינוס לטור טיילור סביב 0:

$$f(x) \approx \frac{1 - \left(1 - \frac{x^2}{2!} + \frac{x^4}{4!} - \frac{x^6}{6!} + \dots\right)}{x^2} = \frac{1}{2!} - \frac{x^2}{4!} + \frac{x^4}{6!} + \dots = \frac{1}{2} - \frac{x^2}{24} + \frac{x^4}{720} + \dots$$

$$\left| \frac{x^6}{8!} \cos(\xi) \right| \quad \text{נותן תוצאה הרבה יותר מדויקת (בידקו שארית לגרנז'}$$

למשל, חישוב עם 6 ספרות בשיטת עיגול ייתן 0 במקרה של חישוב לפי הנוסחה
הנתונה ו-0.5 במקרה של שימוש בטור טיילור. פתרון מדויק 0.499995833347366.



```
>> x=-4*10^(-8):10^(-9):4*10^(-8);  
>> y=(1-cos(x))./x.^2  
>> plot(x,y)
```

מספר מצב (Condition Number)

עבור בעיות ואלגוריתמים

- אחת מהסיבות לדיוק לא טוב בפלט היא הגדרת בעיה לקויה : רגישות הפתרון להפרעות בקלט
- יותר מדויק : הפרעה קטנה בקלט גורמת לשינוי גדול בפלט
- בעיה הרגישה לשינויים קטנים בקלט נקראת **לא יציבה נומרית או מתמטית (ill-conditioned problem/algorithm)**

מספר מצב [Condition Number]

דוגמא לאי יציבות שמקורה בבעיה:
נתבונן במערכת המשוואות

$$\begin{cases} x + y = 2 \\ x + 1.01y = 2.01 \end{cases}$$

שפתרונה הוא $x = y = 1$
וכעת נתבונן במערכת המשוואות

$$\begin{cases} x + y = 2 \\ x + 1.01y = 2.02 \end{cases}$$

שפתרונה הוא $x = 0, y = 2$.

שינוי קטן באגף ימין במאית גורר תשובה שונה לחלוטין

מספר מצב (Condition Number)

- מדד ליציבות הבעיה/ אלגוריתם נקרא **מספר מצב**, מסומן ב- C_P ושווה לשינוי יחסי הכי גדול שיכול להיות בפלט כתוצאה מהפרעה קטנה בפלט

תהי f - בעיה חישובית נומרית $f: \mathbb{R}^p \rightarrow \mathbb{R}^q$, $x \mapsto y = f(x)$. מגדירים נורמה $\| \cdot \|$ על $\mathbb{R}^p, \mathbb{R}^q$. נניח שקלט x - בעל שגיאה יחסית קטנה $\frac{\|\Delta x\|}{\|x\|}$. אנחנו רוצים למצוא חסם עליון

על השגיאה היחסית בפלט $\frac{\|\Delta y\|}{\|y\|}$. אזי מספר מצב מוגדר בצורה הבאה :

$$\frac{\|\Delta y\|}{\|y\|} \leq C_P(y, x) \frac{\|\Delta x\|}{\|x\|}$$

$$\frac{\|\Delta y\|}{\|y\|} \leq C_p(y, x) \frac{\|\Delta x\|}{\|x\|}$$



עבור $f: \mathbb{R} \rightarrow \mathbb{R}$ מהגדרה של מספר מצב נקבל

$$C_p = \max \frac{\left| \frac{f(x + \Delta x) - f(x)}{f(x)} \right|}{\left| \frac{\Delta x}{x} \right|} = \left| \frac{\frac{f'(x)}{f(x)}}{\frac{1}{x}} \right| = \left| \frac{f'(x)}{f(x)} x \right|$$

אם מתקיים כי $C \gg 1$ אזי מצב זה נקרא *ill-conditioned*.
 ככל ש- C_p קטן מ-1, כל הבעיה יציבה יותר.

נורמות

נורמה היא הכללה והרחבה של מושג האורך עבור וקטורים ומטריצות. נסמן את הנורמה של המשתנה x כך: $\|x\|$. נורמה הינה פונקציה שמתאימה לכל וקטור במרחב וקטורי ערך ממשי (או בצורה פורמאלית $\mathbb{R} \rightarrow \vec{x} : \|x\|$). הנורמה מקיימת 3 דרישות בסיסיות שיש לצפות שאורך יקיים:

- אורך הוא תמיד חיובי חוץ מווקטור ה-0 שאורכו 0: $\|x\| = 0 \Leftrightarrow x = 0$
- הכפלת ווקטור בסקלר מכפילה גם את אורכו בסקלר (הכוונה היא לערך המוחלט של הסקלר): $\|c \cdot x\| = |c| \cdot \|x\|$
- בחיבור וקטורים מתקיים אי שוויון המשולש: $\|x + y\| \leq \|x\| + \|y\|$

נורמות שימושיות

נורמה של וקטור

$$\|x\|_t = \sqrt[t]{\sum_{i=1}^n |x_i|^t}$$

נורמה אוקלידית

$$\|x\|_2 = \sqrt{\sum_{i=1}^n |x_i|^2}$$

מרחק מנהטן

$$\|x\|_1 = \sum_{i=1}^n |x_i|$$

נורמה אינסוף

$$\|x\|_\infty = \max |x_i|$$

סכומים

נסמן את הסכום ב- $S = a_1 + a_2 + \dots + a_n = \sum_{j=1}^n a_j$, כאשר כל a_j - מספר בייצוג נקודה צפה.

נגדיר $S_2 = fl(a_1 + a_2)$ - גרסת סכום שני איברים ראשונים בייצוג של נקודה צפה.

בצורה דומה נגדיר $S_3 = fl(a_3 + S_2)$, ..., $S_n = fl(a_n + S_{n-1})$ - גרסת מחשב של S .

נזכר כי $fl(x) = x(1 + \varepsilon)$, ולכן נקבל

$$S_2 = (a_1 + a_2)(1 + \varepsilon_2)$$

$$S_3 = (a_3 + S_2)(1 + \varepsilon_3)$$

...

$$S_n = (a_n + S_{n-1})(1 + \varepsilon_n)$$

נעריך שגיאה :

באידוקציה אפשר להראות כי

$$\begin{aligned} S_n &= (a_1 + a_2)(1 + \varepsilon_2) \dots (1 + \varepsilon_n) + \\ &\quad a_3(1 + \varepsilon_3) \dots (1 + \varepsilon_n) + \\ &\quad a_4(1 + \varepsilon_4) \dots (1 + \varepsilon_n) + \\ &\quad \dots \\ &\quad a_n(1 + \varepsilon_n) \end{aligned}$$

אז

$$\begin{aligned} S - S_n &= (a_1 + a_2)[1 - (1 + \varepsilon_2) \dots (1 + \varepsilon_n)] + \\ &\quad a_3[1 - (1 + \varepsilon_3) \dots (1 + \varepsilon_n)] + \\ &\quad a_4[1 - (1 + \varepsilon_4) \dots (1 + \varepsilon_n)] + \\ &\quad \dots \\ &\quad a_n[1 - (1 + \varepsilon_n)] \end{aligned}$$

אוניברסיטת בר-אילן, תשע"ו. מאת : דר' אלכסנדרה אגרנוביץ'

$$\begin{aligned}
S - S_n &= (a_1 + a_2)[1 - (1 + \varepsilon_2) \dots (1 + \varepsilon_n)] + \\
&\quad a_3[1 - (1 + \varepsilon_3) \dots (1 + \varepsilon_n)] + \\
&\quad a_4[1 - (1 + \varepsilon_4) \dots (1 + \varepsilon_n)] + \\
&\quad \dots \\
&\quad a_n[1 - (1 + \varepsilon_n)]
\end{aligned}$$



$$S - S_n \approx -a_1(\varepsilon_2 + \dots + \varepsilon_n) - a_2(\varepsilon_2 + \dots + \varepsilon_n) - a_3(\varepsilon_3 + \dots + \varepsilon_n) - a_4(\varepsilon_4 + \dots + \varepsilon_n) - \dots - a_n \varepsilon_n$$

מכאן, אפר לראות כי כדי למזער את השיגאה, צריך למיין את איברים הסכום כך שיקיימו

$$|a_1| \leq |a_2| \leq \dots \leq |a_n|$$

כי במקרה כזה הטעות הגדולה ביותר נכפלת במספר הקטן ביותר.