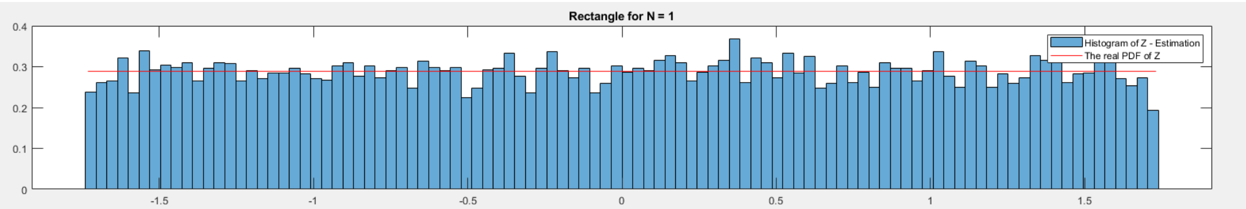
פרויקט סוף – אותות אקראיים ורעש

מגישים: אופיר נחשוני 204616718 , אופק בן עטר 322208430.

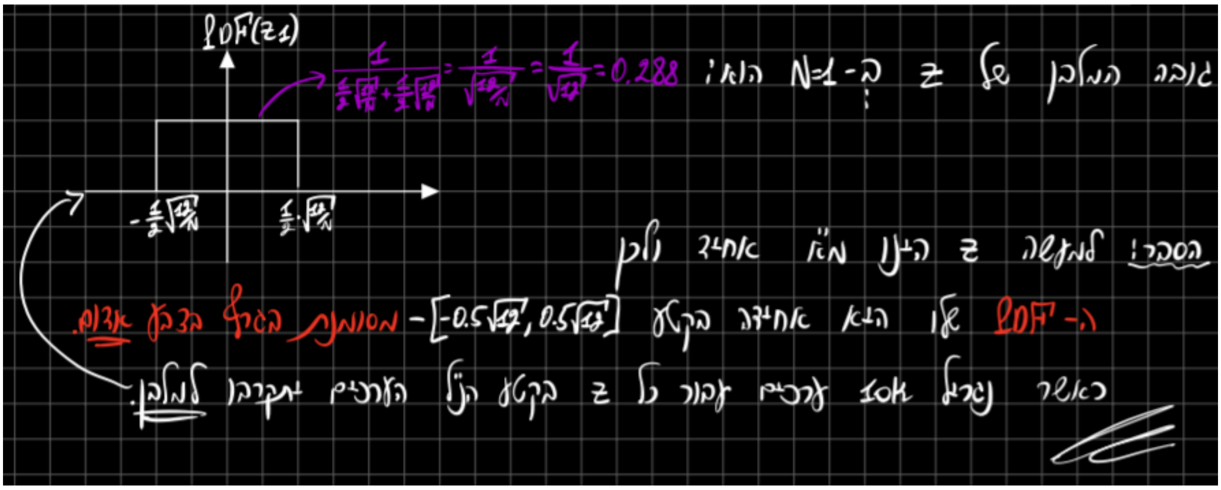
שאלה 1: קבצי מטלב נמצאים בתיקייה Q1

קישור לסרטון הסבר:
https://drive.google.com/uc?export=download&id=1JkJVD2C2MRBxvH5xi0omoqGFPg89nt9c

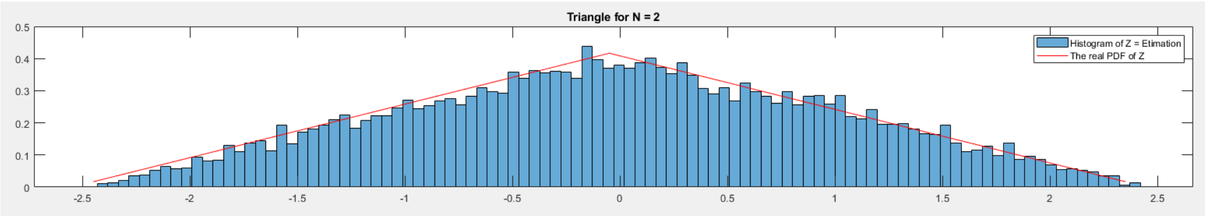
סעיף א' עבור N=1:



הסבר:



<u>המשך שאלה 1 – סעיף ב' עבור N=2:</u>



<u>הסבר:</u>

למה PDF של סכום מ"א בת"ס יוצא הקונבולוציה של ה-PDF-ים?

מכיוון

הגדרת הקונבולוציה

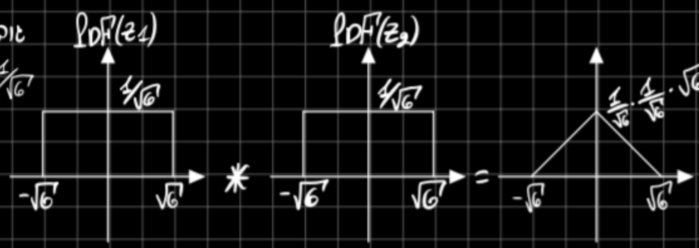
$$f_Z(z_2)=\sqrt{12/2}\,f_Z(x_1+x_2)=\sqrt{6}\int_{\alpha_1+\alpha_2=\alpha} f_{x_1,x_2}(\alpha_1,\alpha_2)\,d\alpha \underset{x_1,x_2\ \text{בת"ס}}{=} \sqrt{6}\int_{-0.5}^{0.5} f_{x_1}(\alpha_1)\,f_{x_2}(\alpha-\alpha_1)\,d\alpha_1$$

PDF של סכום מ"א בת"ס הוא קונבולוציה של ה-PDF-ים: $PDF\{Z\}=\sqrt{12/6}\,(PDF(x_1)*PDF(x_2))$

לכן ה-PDF של המ"א תהיה קונבולוציה בין 2 מלבנים (כל מלבן כמו בסעיף הקודם)

כצפוי, מה שיוצא משולש:

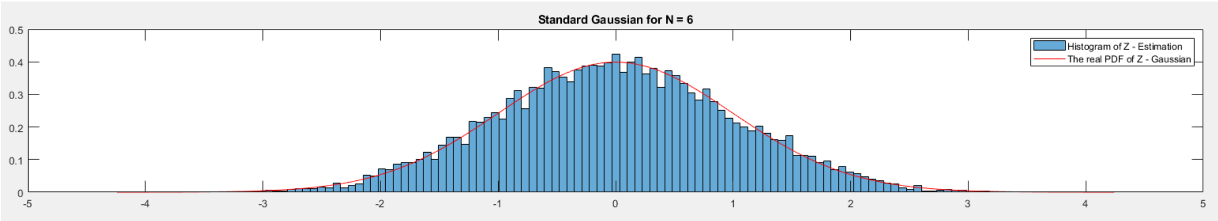
גובה המלבנים: $\frac{1}{\sqrt{12/2}}=1/\sqrt{6}$



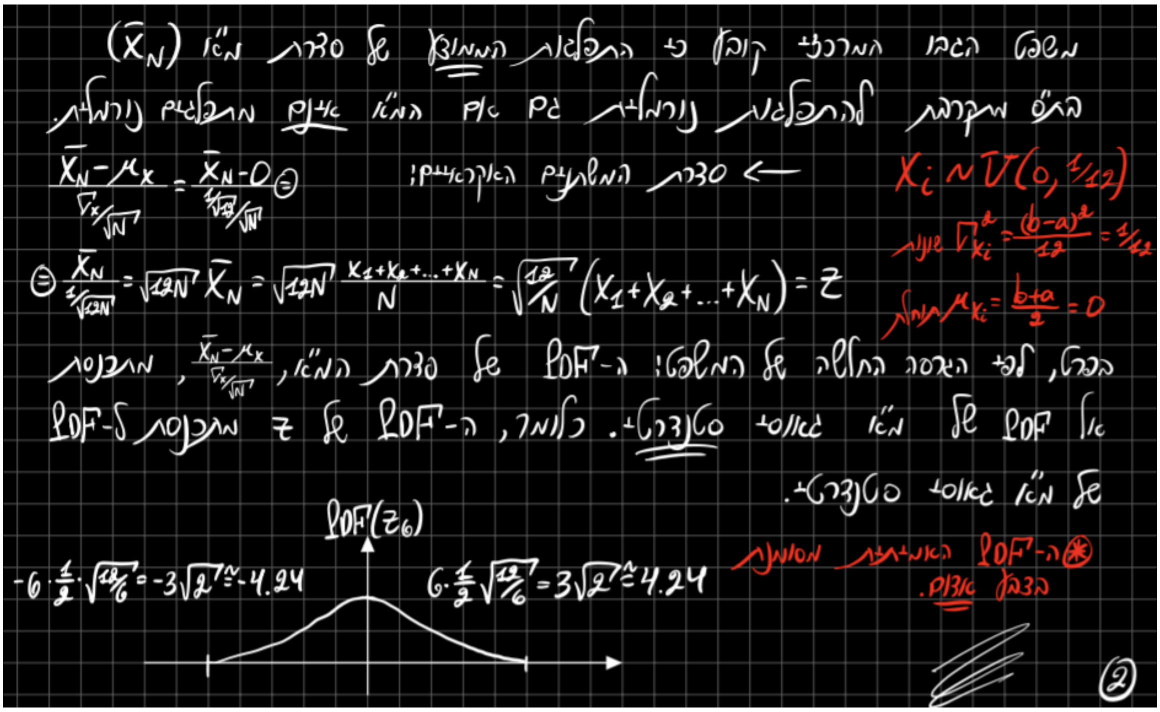
$\frac{1}{\sqrt6}\cdot\frac{1}{\sqrt6}\cdot\sqrt6=\frac{\sqrt6}{6}=0.408$

(*) ה-PDF האמיתית צבועה בצבע <u>אדום</u>.

המשך שאלה 1 – סעיף ג' עבור N=6:



הסבר:



משפט הגבול המרכזי קובע כי התפלגות הממוצע של סדרת מ"מ ($\bar{X}_N$)
בת"ס מקרבת להתפלגות נורמלית גם אם המ"מ אינם מתפלגים נורמלית.

$X_i \sim U(0, 1/12)$

שונות $\sigma^2_{X_i} = \frac{(b-a)^2}{12} = 1/12$

תוחלת $\mu_{X_i} = \frac{b+a}{2} = 0$

← סדרת המשתנים האקראיים:

$$\frac{\bar{X}_N - \mu_X}{\sigma_X/\sqrt{N}} = \frac{\bar{X}_N - 0}{\frac{1}{\sqrt{12}}/\sqrt{N}} \ominus$$

$$\ominus \frac{\bar{X}_N}{1/\sqrt{12N}} = \sqrt{12N}\,\bar{X}_N = \sqrt{12N}\,\frac{X_1+X_2+\ldots+X_N}{N} = \sqrt{\frac{12}{N}}\left(X_1+X_2+\ldots+X_N\right) = Z$$

בפרט, לפי הגרסה החלשה של המשפט: ה-PDF של סדרת המ"מ, $\frac{\bar{X}_N - \mu_X}{\sigma_X/\sqrt{N}}$, מתכנסת אל PDF של מ"מ גאוסי סטנדרטי. כלומר, ה-PDF של Z מתכנסת ל-PDF של מ"מ גאוסי סטנדרטי.

$PDF(Z_6)$

$-6\cdot\frac{1}{2}\sqrt{\frac{12}{6}} = -3\sqrt{2} \cong -4.24$

$6\cdot\frac{1}{2}\sqrt{\frac{12}{6}} = 3\sqrt{2} \cong 4.24$

* ה-PDF האמיתית מסומנת בצבע אדום.

2

שאלה 2: קבצי מטלב נמצאים בתיקייה Q2

קישור לסרטון הסבר:
https://drive.google.com/uc?export=download&id=1qF_Vh2znfYDHe9Qka9I5Wi2uH0ctpZn9

שאלות הכנה אנליטיות (צילומי מסך):

סעיף א':

שאלה 2:

שאלות הכנה אנליטיות:

נתונים:

$\mu_D=0$ D WSS $r_D[k]=\alpha^{|k|}$

$\mu_V=0$ V WSS $r_V[k]=\delta[k]$

$X[n]=D[n]+V[n]$

D – WSS, V – רעש לבן

הוכח: $X[n], D[n]$ WSS במשותף (א)

פתרון:

$V[n], D[n]$ WSS במשותף $\rightarrow r_{V,D}[k]=E(V[n]D[n-k])$

וחס"ק $\rightarrow r_{V,D}[k]=\mu_V\mu_D=0$

נבדוק האם $X[n]$ הוא WSS: (-)

(1) $\mu_x=E(X[n])=E(D[n]+V[n])=E(D[n])+E(V[n])=0+0=0 \rightarrow$ תוחלת של x קבועה

(2) $E(X[n]X[n-k])=E((D[n]+V[n])(D[n-k]+V[n-k]))=E(D[n]D[n-k]+V[n]V[n-k])$ (=)

$E(V[n]D[n-k])=0$

$E(D[n]V[n-k])=0$

$$r_x[k]\overset{?}{=}\text{(=)}\ \underbrace{E(D[n]D[n-k])}_{r_D[k]}+\underbrace{E(V[n]V[n-k])}_{r_V[k]}=\boxed{\alpha^{|k|}+\delta[k]}$$

$\rightarrow$ אוטו-קורלציה לא תלויה ב-n

מכאן ש-$X[n]$ הינו WSS.

נבדוק האם הקרוס קורלציה של $X[n], D[n]$ לא תלויה ב-n: (-)

$V[n], D[n]$ WSS במשותף

וחס"ק! $r_{D,V}[k]=0$

$$r_{x,D}[k]\overset{?}{=}E(X[n]D[n-k])=E((D[n]+V[n])D[n-k])=\underbrace{E(D[n]D[n-k])}_{r_D[k]}+\cancel{E(V[n]D[n-k])}^{0}=r_D[k]=\boxed{\alpha^{|k|}}$$

הקרוס קורלציה לא תלויה ב-n

מכאן שהתהליכים $X[n], D[n]$ הם WSS במשותף!!

המשך שאלה 2 – שאלות הכנה אנליטית:

סעיף ב׳:

שאלה 2:

ב) $\hat{D}[n]=\sum_{k=0}^{p-1}h[k]x[n-k]$

$h[k]$ שמביא את ה-mse למינימום:

$$E(err^2[n])=E\left((D[n]-\hat{D}[n])^2\right)=E\left(\left(\sum_{k=0}^{p-1}h[k]x[n-k]-D[n]\right)^2\right)$$

נגזור לפי $h[m]$ ונשווה ל-0: עבור $m\in\mathbb{Z}$

$$E\left(2x[n-m]\left(\sum_{k=0}^{p-1}h[k]x[n-k]-D[n]\right)\right)=0$$

$$E\left(2x[n-m]\left(\sum_{k=0}^{p-1}h[k]x[n-k]-D[n]\right)\right)=0 \quad /:2$$

$$E\left(\sum_{k=0}^{p-1}h[k]x[n-k]x[n-m]\right)-E\left(x[n-m]D[n]\right)=0$$

$$E\left(\sum_{k=0}^{p-1}h[k]x[n-k]x[n-m]\right)=\underbrace{E\left(x[n-m]D[n]\right)}$$

כי $x[n],D[n]$ WSS במשותף! $\leftarrow$ $r_{x,D}[m]=\alpha^{|m|}$

$$\sum_{k=0}^{p-1}E\left(h[k]x[n-k]x[n-m]\right)=r_{x,D}[m]$$

$$\sum_{k=0}^{p-1}h[k]E\left(x[n-k]x[n-m]\right)=\alpha^{|m|}$$

$$\sum_{k=0}^{p-1}h[k]r_x[m-k]=\alpha^{|m|}=r_{x,D}[k]$$

$k=0,1,\dots,p-1$

למעשה אלו המקדמים של המסנן $h[k]$ שמביא את השגיאה הריבועית הממוצעת למינימום:

$$h[k]=\arg\min\{E(err^2[n])\}$$

$X^T\times h=V$

$(p-1)\times(p-1)$ rows cols, $(p-1)\times 1$ rows cols $=(p-1)\times 1$ rows cols

<u>המשך שאלה 2 – שאלות הכנה אנליטיות סעיף ב׳:</u>

עבור תהליך אקראי ממשי $x[n]$:
$r_x[m-k]=r_x[k-m]$
$r_x[k]\otimes h[k]$

2) ב) $\sum_{k=0}^{p-1} h[k]\,r_x[m-k]=\alpha^{|m|}=r_{x,D}[k]$

$r_x[-k]$ עבור m שלם כלשהו

$$h[k]*r_x[k]=r_{x,D}[k] \qquad r_x[k]*h[k]=r_{x,D}[k]$$

הקונבולוציה הציקלית $r_x[k]$ עם $h[k]$:

⊛ מתכונות קונבולוציה ציקלית - שקילות לכפל בתדר
וגם מתכונות קונבולוציה לינארית - שקילות לכפל בתדר
כלומר קונבולוציה לינארית זהה לקונבולוציה ציקלית.

חישוב מעגלי!

$$h[k]*r_x[k]=h[k]\otimes r_x[k]=\sum_{k=0}^{p-1} h[k]\,r_x[(m-k)\bmod p]=r_{x,D}[k]$$

$m=0,1,\dots,p-1$

⊛ והמשוואה השקולה, בכתיב מטריציוני:

$$\begin{array}{c|c} & k=1 \quad k=2 \quad k=3 \quad k=p \\ \begin{array}{c} m=1\rightarrow \\ m=2\rightarrow \\ m=3\rightarrow \\ \vdots \\ m=p\rightarrow \end{array} & \begin{bmatrix} r_x[0] & r_x[1] & r_x[2] & \cdots & r_x[p-1] \\ r_x[1] & r_x[0] & r_x[1] & \cdots & r_x[p-2] \\ r_x[2] & r_x[1] & r_x[0] & \cdots & r_x[p-3] \\ \vdots & & & & \\ r_x[p-1] & r_x[p-2] & r_x[p-3] & \cdots & r_x[0] \end{bmatrix} \begin{bmatrix} h[0] \\ h[1] \\ h[2] \\ \vdots \\ h[p-1] \end{bmatrix} = \begin{bmatrix} r_{x,D}[0] \\ r_{x,D}[1] \\ r_{x,D}[2] \\ \vdots \\ r_{x,D}[p-1] \end{bmatrix} \end{array}$$

j

<u>משימת מטלב:</u>

<u>סעיף א׳ – פונקציית חישוב המקדמים של המסנן h[k] (optimalFilterMMSE.m):</u>

⊛ משימת matlab: קובץ מטלב בתיקיה Q2

א) פונקציה במטלב - optimalFilteMMSE.m.

optimalFilterMMSE.m

$r_x[k]=\alpha^{|k|}+\delta[k]$

$$\begin{bmatrix} r_x[0] & r_x[1] & \cdots & r_x[p-1] \\ r_x[1] & r_x[0] & & r_x[p-2] \\ \vdots & & \ddots & \vdots \\ r_x[p-1] & r_x[p-2] & \cdots & r_x[0] \end{bmatrix} = \begin{array}{c|c} & j \\ \hline i & \begin{bmatrix} 2 & \alpha & \alpha^2 & \cdots & \alpha^{19} \\ \alpha & 2 & \alpha & & \vdots \\ \alpha^2 & \alpha & 2 & & \\ \vdots & & & \ddots & \\ \alpha^{19} & \alpha^{18} & \alpha^{17} & \cdots & 2 \end{bmatrix} \end{array}$$

$r_x[k]=\alpha^{|k|}+\delta[k]$

$k=i-j$

באלכסון הראשי:

$r_x[1,1]=\alpha^{|1-1|}+\delta[1-1]=\boxed{2}$

$r_{x,D}[k]=r_D[k]=\alpha^{|k|}$

⊛ בכל השאר ה-$\delta[k]$ שווה ל-0!

המשך שאלה 2 – משימת מטלב סעיפים ב' + ג' – סימולציות:

ב) $\alpha=0.8$, $p=1,2,\dots,20$.

• $D[n]$ WSS , $\mu=0$, $r_D[k]=\alpha^{|k|}$ ← לכן, תהליך AR:

0.8

$$D[n]=\alpha D[n-1]+V[n]$$

רעש לבן גאוסי

$\sigma_V^2=?$

$\mu_V=0$

$$\boxed{D[n]=0.8D[n-1]+V[n]}$$

סדרת אוטו קורלציה של D (AR=):
(לפי ההגדרה)

$$r_D[k]=\sigma_V^2\frac{\alpha^{|k|}}{1-\alpha^2}=\alpha^{|k|}$$

$$Var(D[n])=E(D[n]D[n])-\cancel{E^2(D[n])}=r_D[0]=\alpha^0=\boxed{1}=\sigma_D^2$$

$$\boxed{\sigma_V^2}=1-\alpha^2=1-0.8^2=\boxed{0.36}$$

• $V[n]$ הינו רעש לבן WSS ← $\mu_V=0$

$r_V[k]=\delta[k]$

אוטו קורלציה של תהליך WSS $r_V[k]=\sigma_V^2\delta[k]$

$\boxed{\sigma_V^2=1}$

• $X[n]=D[n]+V[n]$ ← חיבור וקטורי

• חישוב השגיאה - MSE: (1) לחשב את המשערך $\hat{D}[n]$:

$p=20$

קונבולוציה לינארית

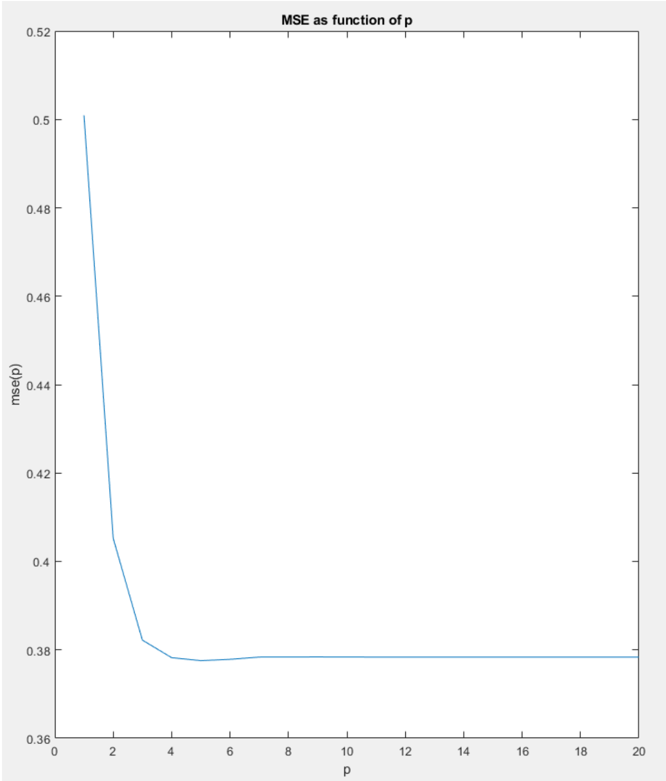
$$\hat{D}[n]=\sum_{k=0}^{p-1}h[k]X[n-k]=\sum_{k=0}^{19}h[k]X[n-k]=h[n]*X[n]$$

(2) לחשב את השגיאה הריבועית הממוצעת:
(מושפעת גם מהמקדם $h[k]$ כי גם הם משוערים)

$$mse=E(err^2)=E\left(\left(D[n]-\hat{D}[n]\right)^2\right)$$

<u>המשך שאלה 2 – סעיף ב׳:</u>

להלן הגרף של שגיאת הMSE כפונקציה של p עבור ערך קבוע של $\alpha = 0.8$:

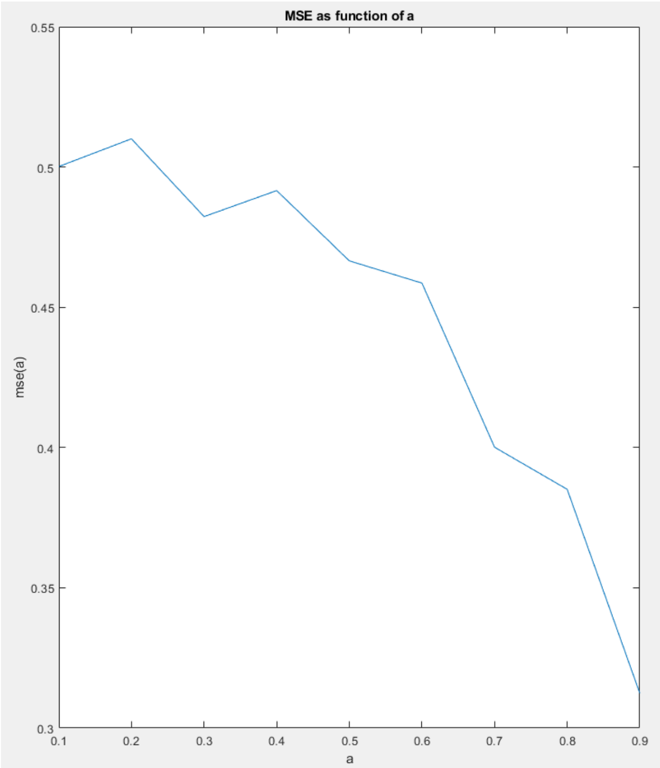


<u>הסבר:</u> בסעיפים האנליטיים הוכחנו שהמקדמים של המסנן – h[k] – הם מביאים למינימום את השגיאה הריבועית הממוצעת. חישבנו את המקדמים של המסנן h[k] שמביא את MSE למינימום, כלומר לMMSE. ככל שישנם יותר מקדמים לh[k] כך המשערך מדיוק יותר, כי אנחנו משערכים על סמך יותר מידע.

<u>ממצאים מהגרף:</u>

ניתן לראות כי הגרף של MSE כפונקציה של p יורד, מכיוון שככל שp גדל כך השגיאה קטנה יותר. ניתן לראות שעבור p=1 השגיאה היא הגדולה ביותר, כיוון שהמסנן כולל רק מקדם אחד. מp=4 השגיאה הריבועית הממוצעת מתייצבת לערך המינימלי שלה. כלומר נדרשים 4 מקדמים של המסנן ע״מ להביא את הMSE לקריטריון של MMSE.

<u>המשך שאלה 2 – סעיף ג':</u>



<u>הסבר:</u> תהליך D[n] הוא תהליך AR אז מהנוסחה שלו עולה כי מכיוון שהוא תלוי בערך עבר שלו. אז אם אלפא גדולה יותר (קטנה מ-1), אז למעשה התהליך מתקרב לערך עבר שלו (עם תוספת של רעש לבן גאוסי). $D[n] = \alpha * D[n-1] + U[n]$

לכן, ככל שאלפא גדלה התהליך D[n] יציב יותר. מכאן שניתן יהיה לשערך את התהליך באופן מדוייק יותר, מה שיקטין את השגיאה הריבועית הממוצעת (MSE).

<u>ממצאים מהגרף:</u>

ניתן לראות שהגרף נמצא <u>במגמת ירידה</u>.

שאלה 3: קבצי מטלב בתיקייה Q3

קישור לסרטון הסבר:
https://drive.google.com/uc?export=download&id=18kal830p3HDe-np-rFDw9AaYjvi21Eum

שאלות הכנה:

1. להלן סיכום המאמר Better Decisions through Science:

הקדמה:

בהרבה תחומים נדרש לבצע דיאגנוזה (בדיקה) ולקבל החלטות מכריעות שמשפיעות על חיי אדם או על תחומים אחרים חשובים. למשל, טכנאי מטוסים בשדה תעופה צריך לקבל החלטה, עפ"י הבדיקות שהוא מבצע, לקבוע האם יש סדק בכנף, כלומר האם המטוס כשיר או לא. יתרה מזאת, רופא שמבצע דיאגנוזה לצילום של שד, צריך לקבל החלטה האם קיים גידול סרטני ממאיר או שפיר. מכאן, ניתן להבין את חשיבות תהליך גיבוש הדיאגנוזה ע"י מקבל ההחלטה בהמון תחומים.

ישנן שיטות/גישות שמשפרות את תהליך קבלת ההחלטה. שיטה מסויימת שמשפרת בתעשייה אחת יכולה להשפיע סימולטנית על תעשיות אחרות.

קיימות לפחות 2 שיטות כאלה. הראשונה משפרת את הדיוק. כלומר, משפרת את הסיכויים שכל החלטה שתתקבל תהיה הנכונה. השנייה משפרת את התועלת של תהליך קבלת ההחלטה. כלומר מבטיחה שמספר המקרים שמתבררים כנכונים (True Positive) לא נובעים ממספר לא סביר של המקרים שמתבררים כשגויים (False Positive).

האנשים שמבצעים את הדיאגנוזה (בדיקה) לא חייבים להיצמד לנוסחאות מתמטיות. בתחומים מסויימים, הסטטיסטיקה (וכלים אובייקטיבים נוספים) משמשים כדעה משנית. כלומר, הכלים הסובייקטיביים של מבצע הבדיקה משמשים כדעה הראשית. בתחומים אחרים, היחס הפוך.

פרק 1(Tools of the Trade):

אם בכל תהליך של אבחון, הכלים הסובייקטיביים יספקו תשובות חד משמעיות נכונות, אז לא יהיה צורך במידע סטטיסטי. במהלך האבחון הכלים הללו אינם מספקים תשובות חד משמעיות. כלומר, נדרש לבצע להם אנליזה – לפרש אותם. לצורך הדוגמא, אם לחץ בתוך העין הוא המדד היחיד לכך שהמטופל חולה במחלת עין, גלאוקומה. אם מכמתים את האופיין, אז נניח שמתחת ל10 האדם בריא ומעל 40 האדם חולה, אך מה קורה בתחום שבין 10 ל-40. יש לפרש את התוצאה (ניתן לעשות זאת באמצעות כלים סטטיסטיים). ע"מ לפתור את הבעיה הזו, יש לקחת מדגם של מטופלים בעלי תוצאות של לחץ על העין. יש לאחר מכן לבצע מחקר על דגימות נתוני הלחץ של אוכלוסיית החולים לעומת הבריאים, ולחפש את המדד שקובע שמעליו (או מתחתיו) המטופל חולה או בריא.

האלגוריתמים המתמטיים שמפרטים את המבחנים הטובים ביותר שכלולים בדיאגנוזה ומחשבים את הסבירות, מבוססים על שילוב של תוצאות ממדידות וידועים כאמיתות סטטיסטיות (SPR).

הנתונים האובייקטיבים (כמו הערך של הלחץ בעין), מן הסתם, משמש כדי לשפר את הדיוק בחיזוי הסטטיסטי. יתרה מזאת, החיזוי הסטטיסטי יכול להתבצע גם עם עבודה של נתונים סובייקטיביים, באמצעות רשימה מפורשת של קריטריונים תפיסתיים שיכולים להיות מדורגים ספקטרלית מ-1 עד 5 למשל.

לצורך חיזוי אופטימלי, על מקבלי ההחלטות לקבוע אילו חוקים סטטיסטיים, שמשמשים לחיזוי, יתנו דיוק רב יותר. ע"מ לקבוע את הדיוק הכולל של כלי חיזוי אלה, משתמשים בעקומות ROC (Receiver Operating Characteristic). דוגמא לשימוש בעקומות הROC התרחשה במלחמת העולם השנייה, כאשר ציוד של רדאר מסויים מבדיל בין אותות של שמייצגים רעש או הפרעה כלשהי לבין אותות שמקורם ממטוסי אויב.

תוכנות שממשות את עקומות הROC נדרשות לקבוע איך יסווגו התוצאות ביחס לעקומה שנבחרה. אופן הסיווג נעשה באמצעות קביעת ערך הסף (threshold). נשאלות השאלות הבאות (ניקח את דוגמת החולים בגלאוקומה):

<u>המשך שאלה 3 – שאלות הכנה סעיף 1:</u>

א. כמה אחוזים מבין אלו שנחזו כחולים באמת חולים במחלה. נקרא True Positive Decisions.

ב. כמה אחוזים מבין אלו שנחזו כחולים לא באמת היו חולים במחלה. נקרא False Positive Decisions.

התוכנות מחשבות את אחוזים אלה. עבור כל ערך סף (threshold) משרטטת גרף של האחוזים כפונקציה של הthreshold. הthreshold ינוע בתחום 0%-100%. הגרף יעלה ככל שthreshold גדל. המסקנה הנוספת היא שתלילות הגרף קובעת את דיוק החוק הסטטיסטי שנעשה בו שימוש. כלומר החיזוי יהיה מדוייק יותר, מכיוון שמספר המשתתפים שנחזו נכון שהם חולים גדול יותר ביחס למספר המשתתפים שנחזו לא נכון שהם חולים.

ערך הסף, הthreshold, הינו עדין ויש לקבוע בתבונה, כך שיביא לניבוי מקסימלי. כלומר ערך סף שיביא לשגיאה מינימלית בחיזוי האם המחלה קיימת אצל המשתתף.

למרבה המזל, ישנם מספר כללי אצבע וכלים מתמטיים שיעזרו לקבוע את ערך הסף.

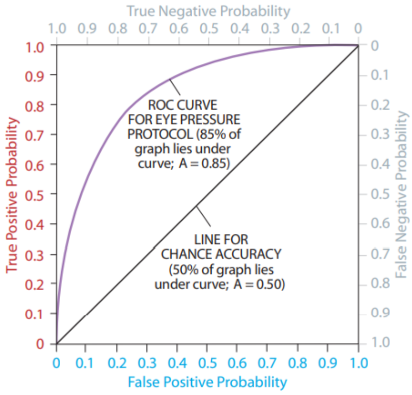
<u>פרק 2 (Rules come to Life) ופרק 3 (Better Cancer Diagnoses):</u>

המודל שמשתמש בעקומות הROC הינו שכיח ביותר בתחומים שבהם נדרשת תשובה לשאלה של כן או לא. באמצעות הדוגמא של מדידות לחץ הנוזלים בעין נוכל להסביר זאת באופן פשטני. אופן השימוש בשיטה זו מתחלקת לארבעה שלבים מרכזיים:

שלב ראשון – בחירת האוכלוסייה המשתתפת במחקר (במדידות), כאשר ידוע המידע האם המשתתפים חולים בגלאוקומה או לא וידועה מידת לחץ הנוזלים בעין שלהם. חלוקת הקבוצה המשתתפת ל-2 קבוצות: הראשונה חולים בגלאוקומה (עקומה אדומה) והשנייה בריאים (עקומה כחולה). שרטוט גרף של כמות המשתפים כפונקציה של לחץ הנוזלים בעין.

שלב שני – חישוב ההסתברות לחיזוי נכון לכך שהמשתתף חולה (True Positive) וחישוב ההסתברות לחיזוי שגוי שהמשתתף חולה (False Positive), באמצעות השוואת מספר המקרים לthreshold (ערך הסף).

שלב שלישי – שרטוט גרף של TP כפונקציה של הFP. הקו הלינארי הנמתח בדיוק באמצע מייצג את הסיכוי לניבוי של 50/50 (הקו השחור). לעומת זאת ככל שהעקומה מתפתלת לכיוון שמאל (הקו הסגול) כך החיזוי טוב יותר. A מייצג את מידת הדיוק של חוק החיזוי הסטטיסטי שנבחר.



שלב רביעי – ככל שעקומת הדיוק (A) הסגולה מתפתלת יותר לכיוון שמאל כך הדיוק נכון יותר. מכאן יש לבחור את ערך הסף האופטימלי (הthreshold) – כלומר יחס מינימלי של הTP לעומת הFP.

<u>המשך שאלה 3 – שאלות הכנה סעיף 1:</u>

<u>פרק 4 (Good Chance of Rain):</u>

דוגמא חשובה היא בתחום חיזוי מז"א מספרת על שימוש במאפיינים אובייקטיבים שנקבעים ע"י מי שתכנן את תוכנת המחשב. תוכנת המחשב מייצרת פלט של גרף עקומות ROC ע"ב נתונים סטטיסטיים. את התוכנה שולחים לחזאי מז"א מוסמכים ומנוסים והם מוסיפים לתוכנה כלים סטטיסטיים סובייקטיביים ראויים לשקלול בחיזוי של התוכנה.

<u>פרק 5 (Thorny Thresholds):</u>

שיטה זו של סיווג ההסתברויות וחיזוי לפי עקומות הROC השפיעה באופן משמעותי ביותר ומכריע על אופן הבדיקות שנעשות להיתכנות איידס (HIV). שיטה זו הפכה ליעילה ביותר לרופאים לקבוע מי נושא את הווירוס איידס, בכך שהחליפה שאת השאלה מי נגוע ומי לא, לשאלה מי בכלל צריך לבדוק. התובנה המרכזית בתהליך זה, שהתפתח בבדיקות האיידס, היא שעדיף להגדיל את הסיכוי של לקבל True Positive, גם אם מחיר הספיגה של False Positive הוא גבוה. תובנה נוספת חשובה ביותר היא יכולת המשחק עם ערך הthreshold ביחס לאוכלוסייה שהרופא מאבחן.

<u>פרק 6 ואחרון (A Plea):</u>

למרות כוחם העצום והשפעתם המכרעת של הכלים הסטטיסטיים לחזות את הערך האמיתי או לסווג בצורה אופטימלית את הנתונים שהמחשב מקבל, ישנן סיטואציות של קבלת החלטות, שלא בהכרח נרצה שמספר המקרים של False Positive(FP) יהיו גדולים/רבים. למשל, ניקח סיטואציה שבה הטייס מקבל את החיוויים ממערכות המטוס שמתריעות על תקלה כלשהי במערכות המטוס. אם מספר המקרים שבהם החיוויים יהיו שגויים (כלומר FP), אז הטייס עלול להתייחס להיות לא מספיק דרוך לתפעל מקרה אמת של תקלה במערכת של המטוס.

<u>המשך שאלה 3 – שאלות הכנה:</u>

2. סיכום הסעיפים 1-3 ממאמר An introduction to ROC analysis:

<u>סעיף ראשון – הקדמה:</u>

גרף הROC זה טכניקה שמאפשרת לנו לבחור איזה מסווג יהיה האופטימלי, ע״ב ביצועיו. מטרתו העיקרית של הROC היא לתאר את הפשרה שנעשית בין TP (True Positive) לבין FP (False Positive) של מסווגים. נעשה שימוש נפוץ בגרף הROC בעיקר במערכות שמייצרות סיווג של כן או לא. בסיטואציות של קבלת החלטה חד משמעית: חולה או בריא, 0 או 1 וכד׳. נייג׳ל ספקמן, אחד ממאמצי שיטת הסיווג של ROC, השפיע רבות והעלה את השימוש בה באופן חד למגוון רחב של שימושים. למרות זאת, ישנם מתנגדים אשר מעריכים כי שיטה זו היא במספר רב של המקרים מדד גרוע עבור מדידת ביצועים.

שיטת הROC מבחינה קונספטואלית היא פשוטה, אך יש לה מורכבויות רבות כאשר משתמשים בה בביצוע מחקרים. מאמר זה, מנסה להקנות את הבסיס של שיטת הגרפים ROC וכמדריך לשימוש בהם במהלך ביצוע מחקר. מטרת המאמר היא להעמיק את הידע אודות גרפים מסוג ROC ע״מ לשפר בעתיד את שיטות הסיווג וההערכה של נתונים.

<u>סעיף שני – ביצועי מסווג:</u>

סוגים שונים של מסווגים שאפשריים הם:

א. סיווגים שנותנים פלט רציף כלשהו. למשל, הפרדה בהסתברות שאובייקט מסויים יהיה משוייך לאחת מ-2 מחלקות.

ב. סיווגים שנותנים פלט בדיד כלשהו. למשל, הפרדה מהו הערך הצפוי ומסווגת את האובייקט לאחת מ-2 המחלקות הנתונות.

בהינתן מסווג ומופע מסויים, ישנן 4 תוצאות אפשריות שונות לפלט הנעשה בתהליך הסיווג:

1. אם המופע חיובי ומסווג כחיובי ע״י המסווג – True Positive.

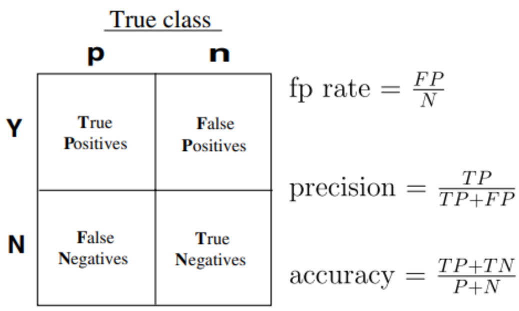
2. אם המופע חיובי ומסווג כשלילי ע״י המסווג – False Negative.

3. אם המופע שלילי ומסווג כחיובי ע״י המסווג – False Positive.

4. אם המופע שלילי ומסווג כשלילי ע״י המסווג – True Negative.

הפלט הזה נהוג להיות מיוצג בצורת מטריצה הנקראת Confusion Matrix.

כפי שרואים באיור Fig1:



באלכסון הראשי של המטריצה נמצאות כל ההחלטות הנכונות (מציאותית). ובאלכסון המשני מופיעות ההחלטות השגויות גם לחיובי וגם לשלילי. כלומר הסטיות מערך האמת.

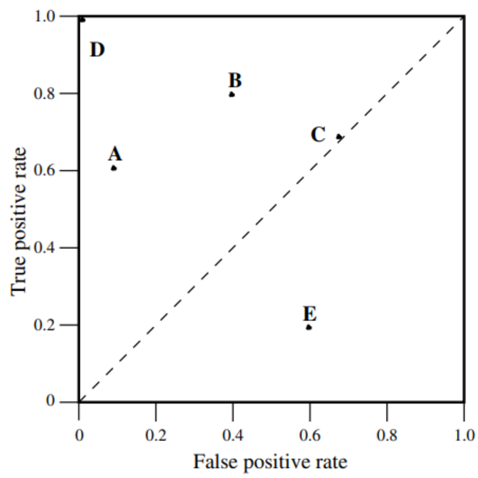
<u>המשך שאלה 3 – שאלות הכנה סעיף 2:</u>

<u>סעיף שלישי –מרחב הROC:</u>

גרף ROC משורטט כאשר הציר האנכי שלו הוא שיעור הTP כפונקציה של שיעור הFP.

איור Fig 2 מתאר גרף ROC של 5 מסווגים שונים המסומנים כA עד E. מסווגים בדידים הם מסווגים שמוציאים פלט בדיד של מספר מחלקה אליה משוייך המופע. כל מספר כזה משוייך לאיבר בטבלה (=Confusion Matrix) מה שמייצג צמד ערכים לזוג (FP,TP) וממוקם על גרף הROC. באיור זה מוצגים מסווגים בדידים בלבד.

להלן הגרף של חמישה המסווגים הבדידים מ-Fig 2:



ישנן כמה נקודות חשובות להתייחסות בגרף זה:

ראשית הצירים (0,0) – הנקודה מייצגת את האסטרטגיה שבה המסווג מחד, לא מבצע שגיאות, כי אין לו FP, ומאידך לא מייצר שום הצלחות, כי אין לו גם TP.

הנקודה (0,1) – מייצגת את הסיווג המושלם, משתי סיבות: אחת מכיוון ש-100% מהמקרים הם הצלחות, כולם TP, והשנייה כי אין למסווג כישלונות, אין לו בכלך FP. כמו למשל המסווג D.

הנקודה (1,1) – מייצגת את הסיווג המקסימלי מבחינת הצלחות, ומאידך מספר הכישלונות המרבי.

באופן כללי, לא פורמלי, אם נקודה מתקרבת יותר <u>לצד העליון שמאלי</u> של הגרף אז היא טובה יותר מנקודה אחרת מנקודה שנמצאת דרום מערבה ביחס אליה (צפון = למעלה).

נקודה A 'קונסרבטיבית' יותר מנקודה B במובן הזה שיש פחות כישלונות (FP), אבל היא מייצרת מספר מועט של הצלחות (TP).

נקודה 'ליברלית' במובן הזה שאם היא נמצאת בצד הימני עליון של הגרף, שיש למסווג הצלחות בכמות גבוהה (TP), אבל היא גורמת למספר רב של כישלונות (FP).

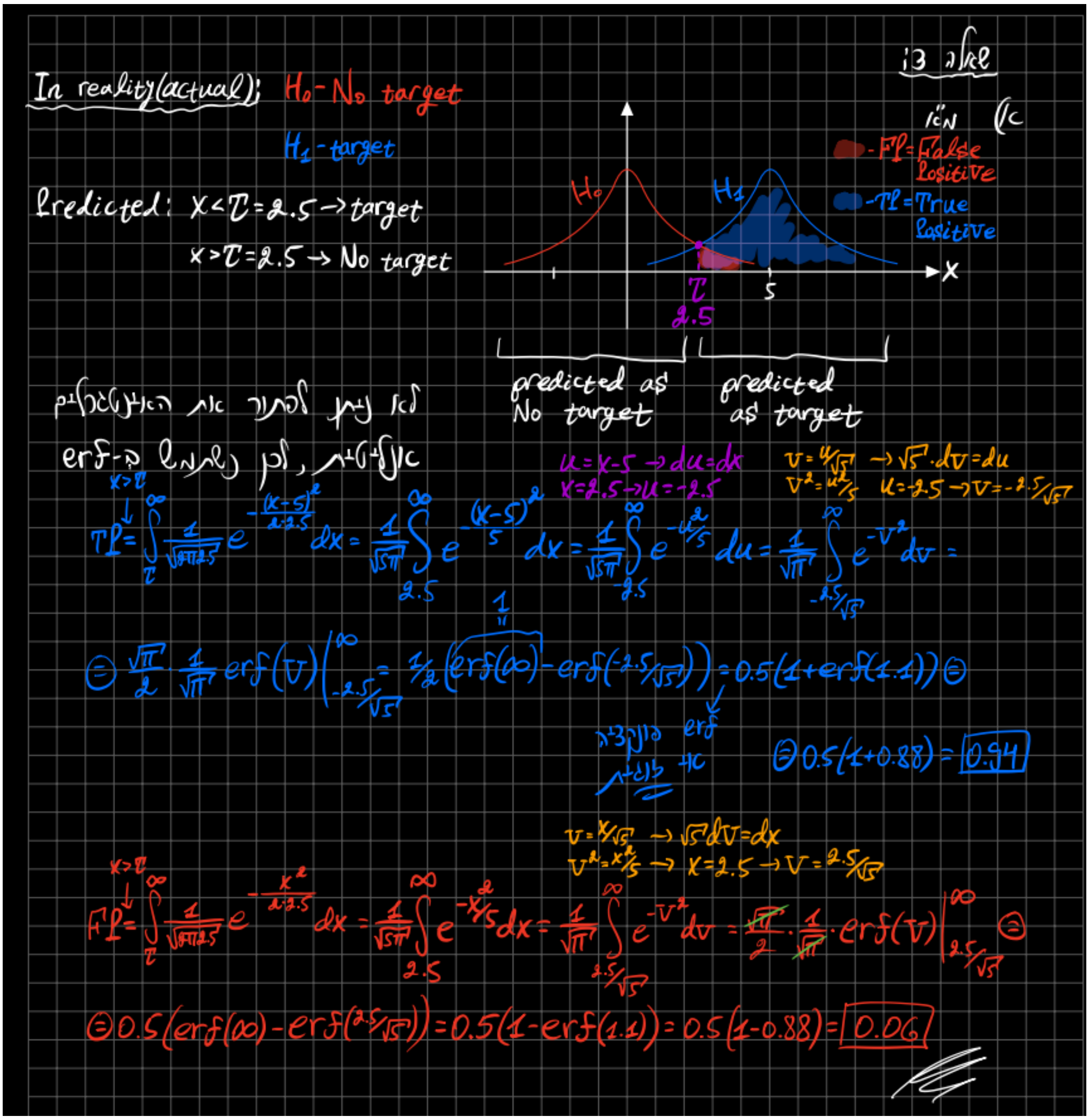
הקו המקווקו מייצג את הסיווג ה'רנדומאלי' לחלוטין. במצב שבו המסווג רנדומלית מנחש על הגרף לאיזה מחלקה ישויך המופע. כלומר אם יוגרלו מספר רנדומלית כחיובי או שלילי, אז חצי מהחיוביים יסווגו כחיוביים (TP) וחצי מהשליליים יסווגו כשליליים (TN).

המשמעות של זה, שאם מסווג כזה מסווג 90% ממחלקת החיוביים (TP) נכון, זה אומר שגם הסיווג הלא נכון של החיוביים הוא מסווג כ90% (FP). כלומר על גרף הROC המסווג ימוקם בנקודה (0.9,0.9). מסווג אשר יש לו אחוז גבוה של הצלחות, אבל אחוז גבוה גם כן בכמות הכישלונות של ניבוי מספר חיובי בהינתן מספר שלילי (FP).

המשך שאלה 3 – משימת matlab:

סעיף א':

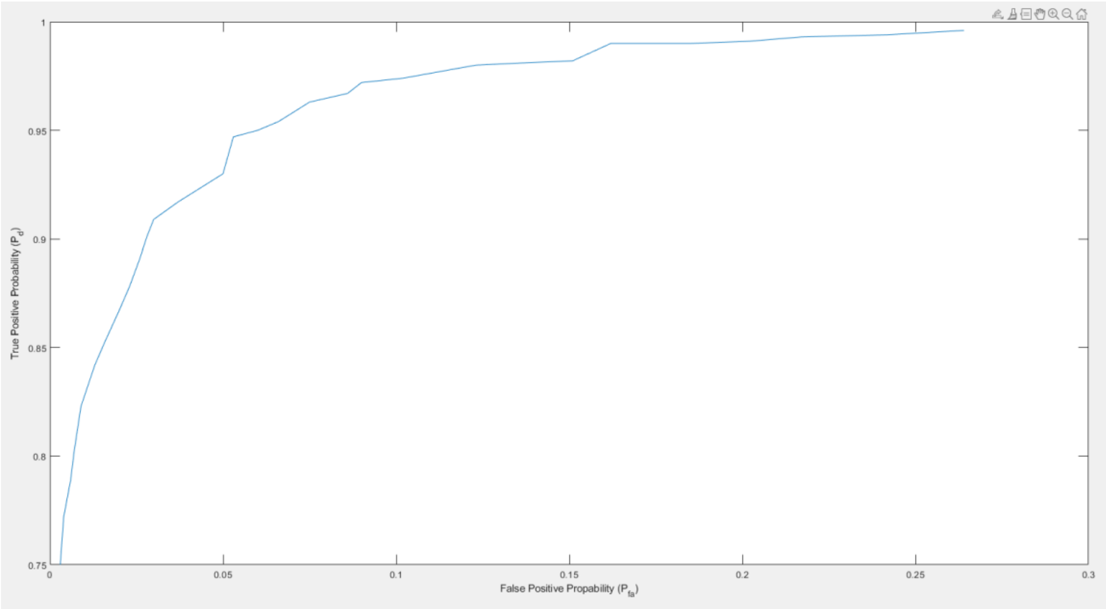
להלן החישוב האנליטי שביצענו עבור TP (True Positive) ו-FP (False Positive):



	תוצאות חישוב אנליטי	תוצאות סימולציית matlab
TP	0.94	0.933
FP	0.06	0.063

ניתן לראות כי השגיאה בין התוצאות קטנה יחסית.

<u>המשך שאלה 3 – משימת מטלב סעיף ב':</u>



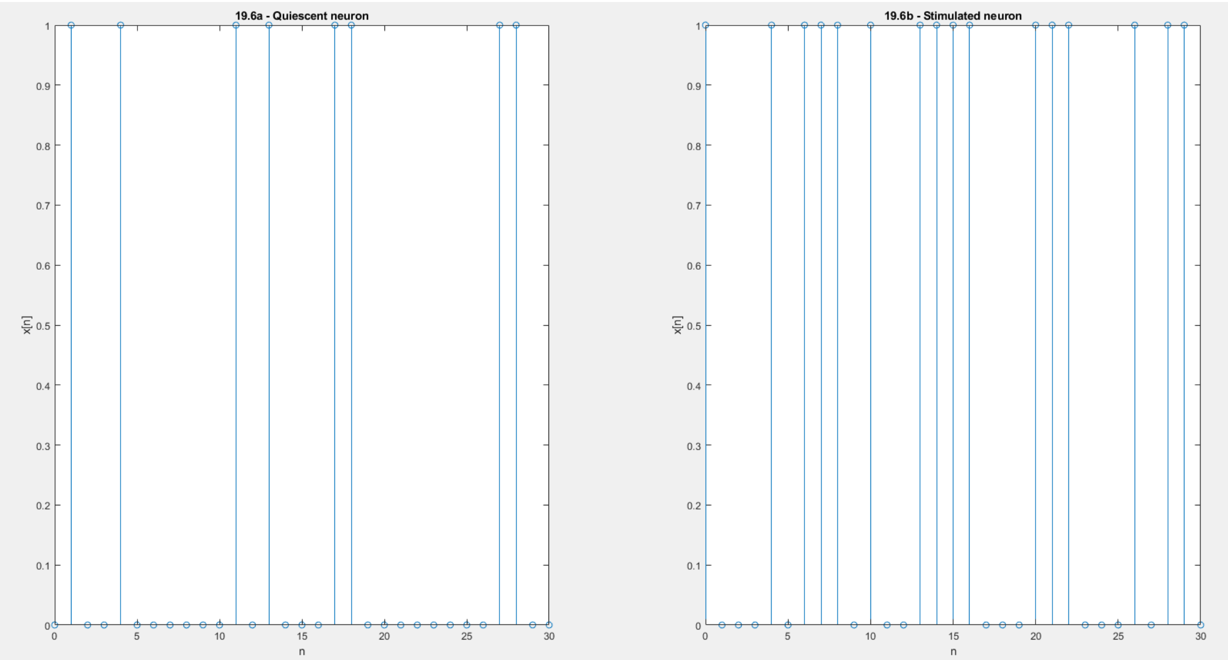
<u>הסבר:</u> למעשה העקומה שאנו רואים היא עקומת הROC והיא מייצגת את הדיוק. כל נקודה על הגרף מייצגת מסנן אחר, כלומר Threshold שונה. מקיעור העקומה ניתן להסיק כי ערכי הסף המסננים האפשריים שיצרנו הינם טובים יותר. לכן, הסינונים יהיו טובים מאשר הקו הלינארי שעובר באמצע הגרף.

שאלה 4: קבצי מטלב בתיקייה Q4

קישור לסרטון הסבר:
https://drive.google.com/uc?export=download&id=1v1dZaKIY48QC-Pl87eIoIMqS-fyKIkg1

להלן כלל הגרפים שהופיעו בסעיף 19.8 מהמאמר שנמצא בקובץ "עזר לשאלה 4".

גרפים של Fig 19.6:



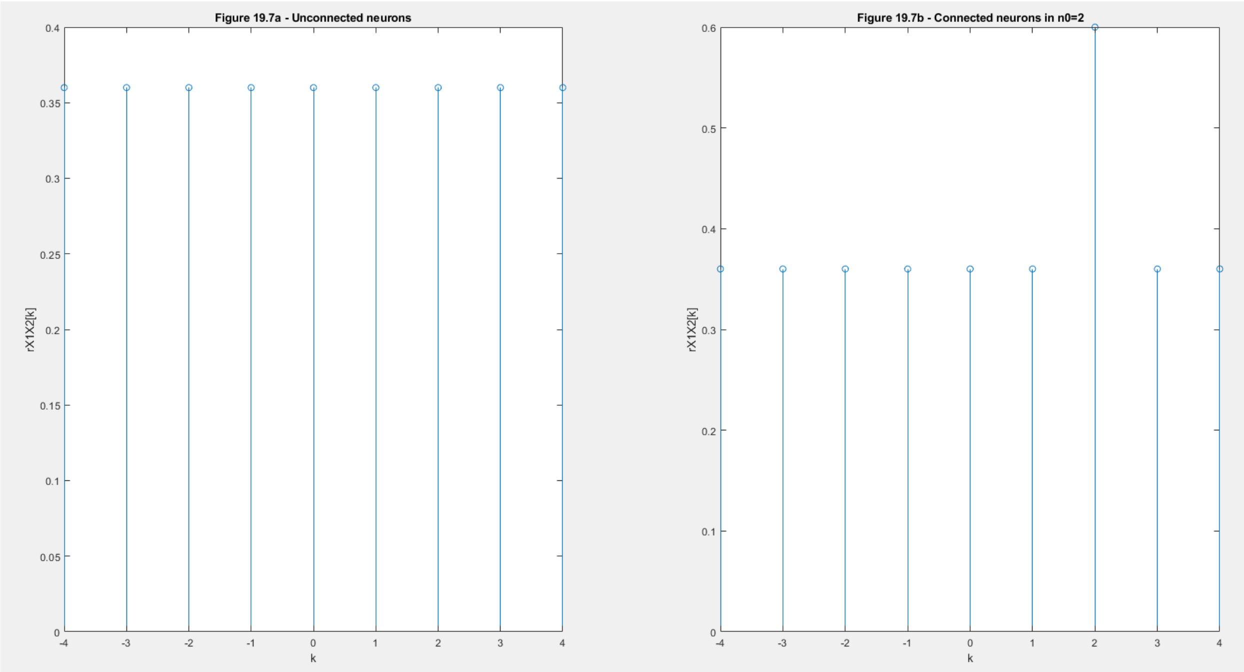
הסבר: אנו רואים שתי רשתות נוירוניות אחת מהן שקטה (quiescent) והשנייה מגורה (stimulated). כלומר, מימשנו את שתי הרשתות הללו בתור תהליכים אקראיים ברנולי (תהליך IID), שבכל זמן n ישנו מ"א ברנולי.

ברשת השקטה (גרף השמאלי) ההסתברות לקבל 1 היא $p_s = 0.1$.

וברשת המגורה (גרף ימני) ההסתברות לקבל 1 היא $p_s = 0.6$.
כאשר נגריל אותם 31 פעמים עבור תהליך אקראי מסויים, נקבל את הגרפים הנ"ל.

<u>המשך שאלה 4:</u>

<u>גרפים של Fig 19.7:</u>



<u>הסבר:</u> הגרפים שבFigure הנ"ל מתארים את פונקציית הקרוס קורלציה של התהליכים האקראיים X1,X2. אלה הפונקציות התיאורטיות של הקרוס קורלציה (CCS). בגרף אחד (הימני) אנו רואים שתי רשתות <u>"מחוברות"</u> ובגרף השני (שמאלי) אנו רואים שתי רשתות <u>לא מחוברות</u>.

רשתות מחוברות מוגדרות כשרשתות אשר האחת היא הזזה בזמן של השנייה ב-n_0, כלומר:

$$X_2[n] = X_1[n - n_0]$$

כאשר $n_0 = 2$, $p_s = 0.6$ (s = state of the neuron)

פונקציית הקרוס קורלציה של רשתות נוירונים X1,X2 מחוברות היא:

$$r_{X1,X2}[k] = r_{X1}[k - n_0] = p_s(1 - p_s)\delta[k - n_0] + p_s^2$$

מהביטוי הזה מגיע הערך המקסימלי בk=2 בגרף הימני, כאשר הרשתות מחוברות.

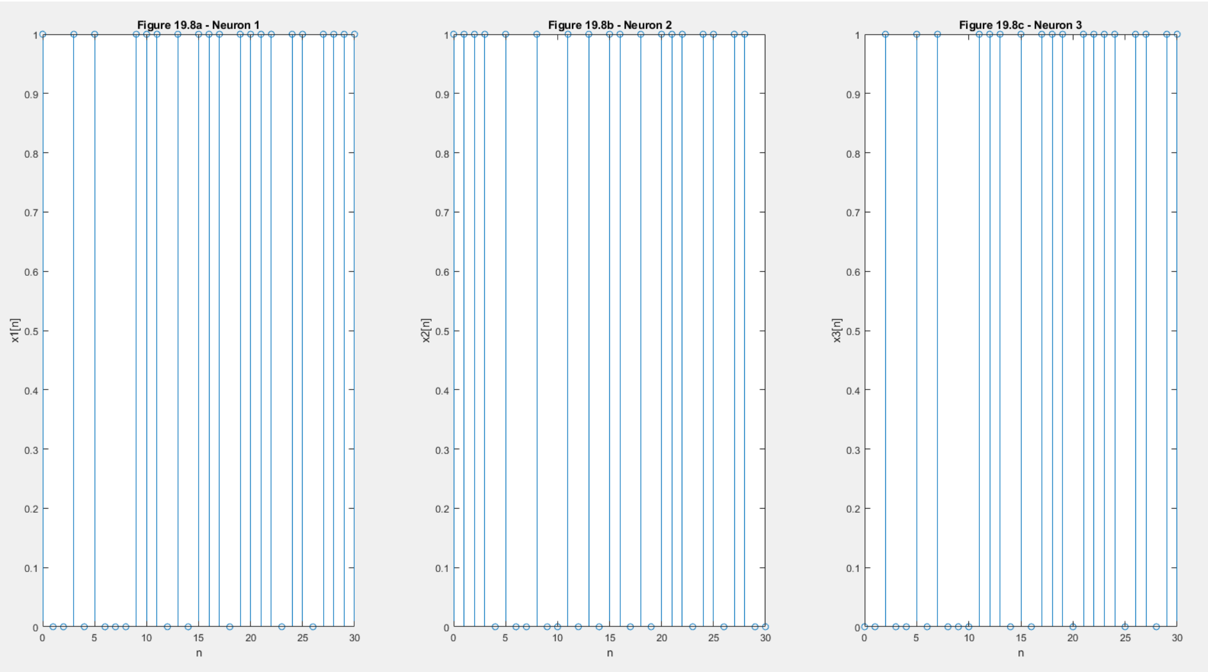
וברשתות נוירונים שלא מחוברות לא תהיה קורלציה ביניהן. לכן, פונקציית הקרוס קורלציה שלהן תהיה:

$$r_{X1,X2}[k] = p_s^2$$

לכן, בגרף הימני פונקציית האוטו קורלציה שווה ל0.36.

<u>המשך שאלה 4:</u>

<u>גרפים של Fig 19.8:</u>



<u>הסבר:</u> הגרפים שבFigure הנ"ל מתארים 3 רשתות נוירונים מגורים (stimulated) X1,X2,X3. כלומר ההסתברות לקבל 1 עבור כל מ"א בזמן n נתון הוא $p_s = 0.6$.

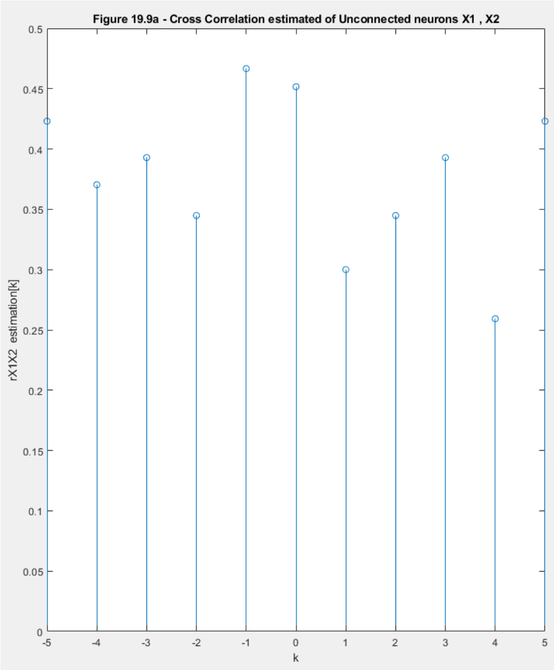
כמו כן, רשתות הנוירונים X1,X3 הן למעשה מחוברות, מכיוון שX3 היא למעשה הזזה בזמן בn_0 של X1. כלומר:

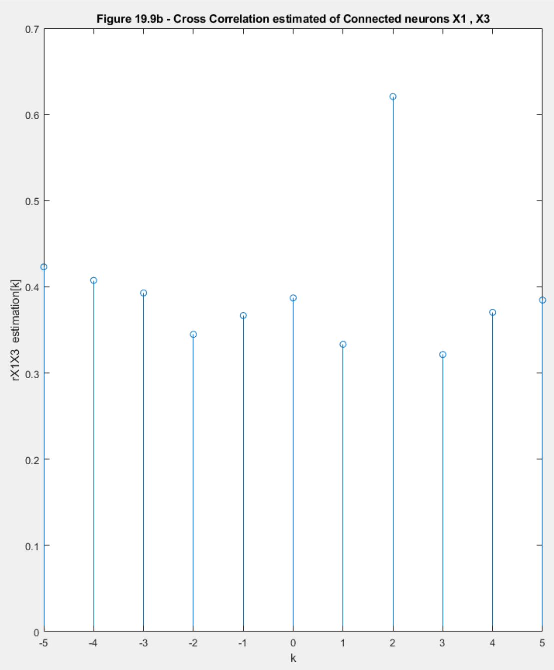
$$X_3[n] = X_1[n - n_0] = X_1[n - 2]$$

לעומת זאת, הרשתות X1,X2 לא מחוברות. נראה את המשמעות בשוני של הקרוס קורלציה של הזוגות הללו בסעיף הבא.

<u>המשך שאלה 4:</u>

<u>גרפים של Fig 19.9:</u>





<u>הסבר:</u> הגרפים שמוצגים בfigure הנ"ל מתארים את פונקציית השערוך של הקרוס קורלציה בין שתי רשתות נוירונים (תהליכים אקראיים למעשה – בדידים). בגרף אחד (הימני) אנו רואים שערוך של פונקציית קרוס קורלציה בין שתי רשתות <u>מחוברות</u> ובגרף שני (שמאלי) אנו רואים שערוך של פונקציית קרוס קורלציה בין שתי רשתות <u>לא מחוברות</u>.

הגרף הימני ($r_{X1,X3}[k]\_estimation$) מכיל ערך מקסימום בk=2 כפי שמופיע בסעיף ב'.

השוני הוא שכאן אנו רואים את המשערך של פונקציות הקרוס קורלציה של שתי רשתות X1,X3 <u>מחוברות</u> ושל שתי רשתות X1,X2 <u>לא מחוברות</u>.

פונקציית המשערך של הקרוס קורלציה היא:

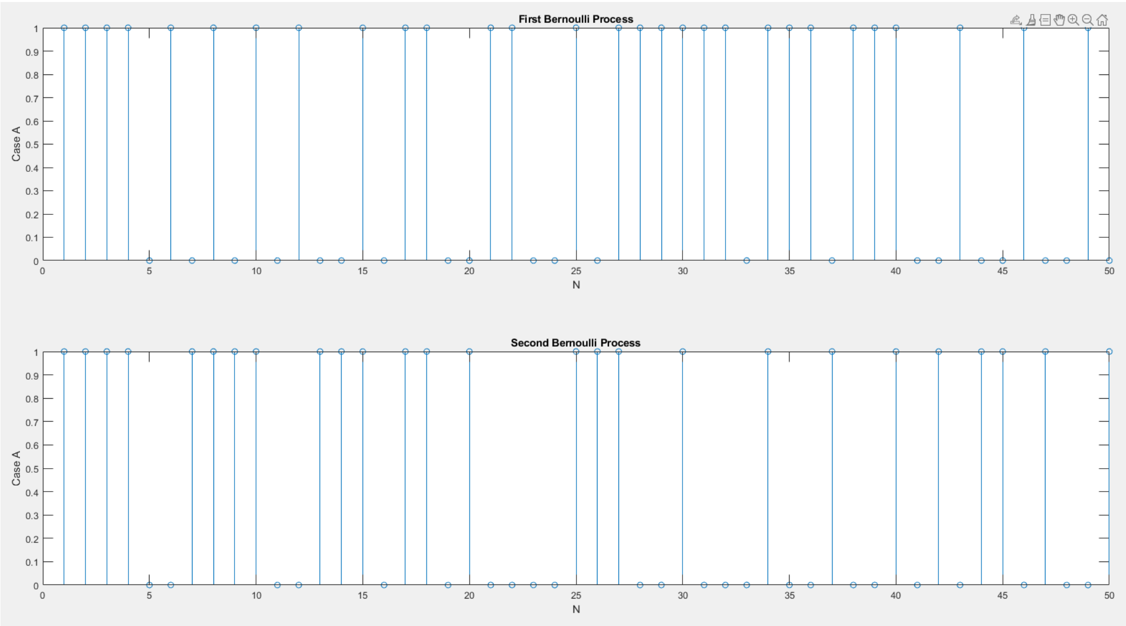
$$r_{X1,X2_{estimation}} = \left\{ \begin{array}{c} \frac{1}{N-k}\sum_{n=0}^{N-1-k} X_1[n] * X_2[n+k], k = \ 0, 1, \dots, M \\ \frac{1}{N-|k|}\sum_{n=|k|}^{N-1} X_1[n] * X_2[n+k], k = \ -M, -(M-1), \dots, -1 \end{array} \right\}$$

כאשר: $N = 31\ , M = 15\ , k = \ -5:5$

<u>שאלה 5:</u> קבצי מטלב בתיקייה Q5

<u>קישור לסרטון הסבר:</u>
https://drive.google.com/uc?export=download&id=1KDR64OSk_daM83mxaJeK9-pIikQ-Jlv0

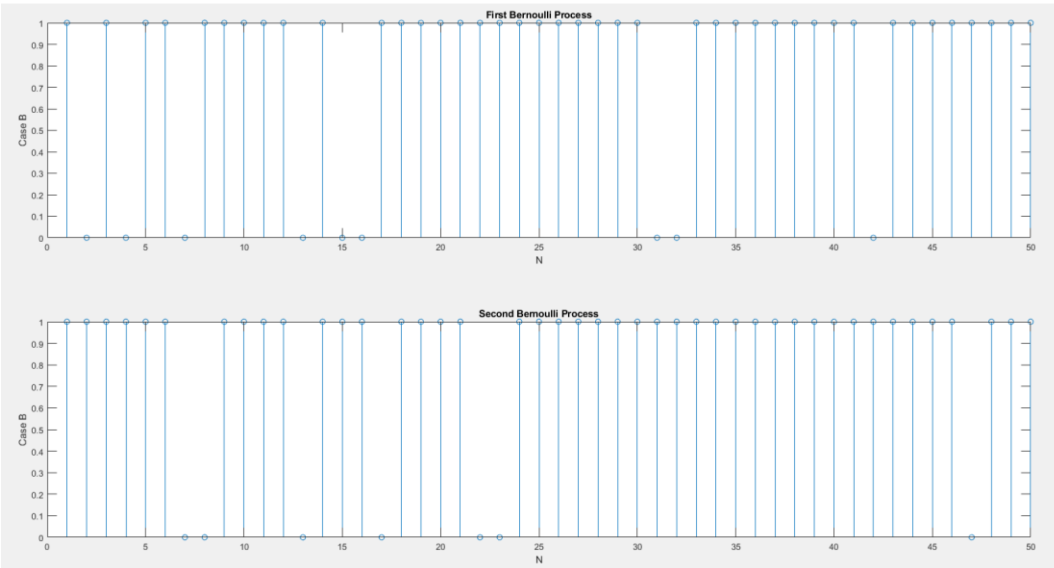
<u>סעיף 1:</u>



<u>הסבר:</u> למעשה בFigure הנ"ל מתוארים תהליכים אקראיים ברנולי, אשר לכל מ"א מסויים יש הסתברות של 0.5 להיות 1 והסתברות של 0.5 להיות 0.

ניתן לראות כי יש שוני בין הגרפים, מכיוון שאנחנו מגרילים 50 מ"א לכל תהליך. השוני נובע מהאקראיות.

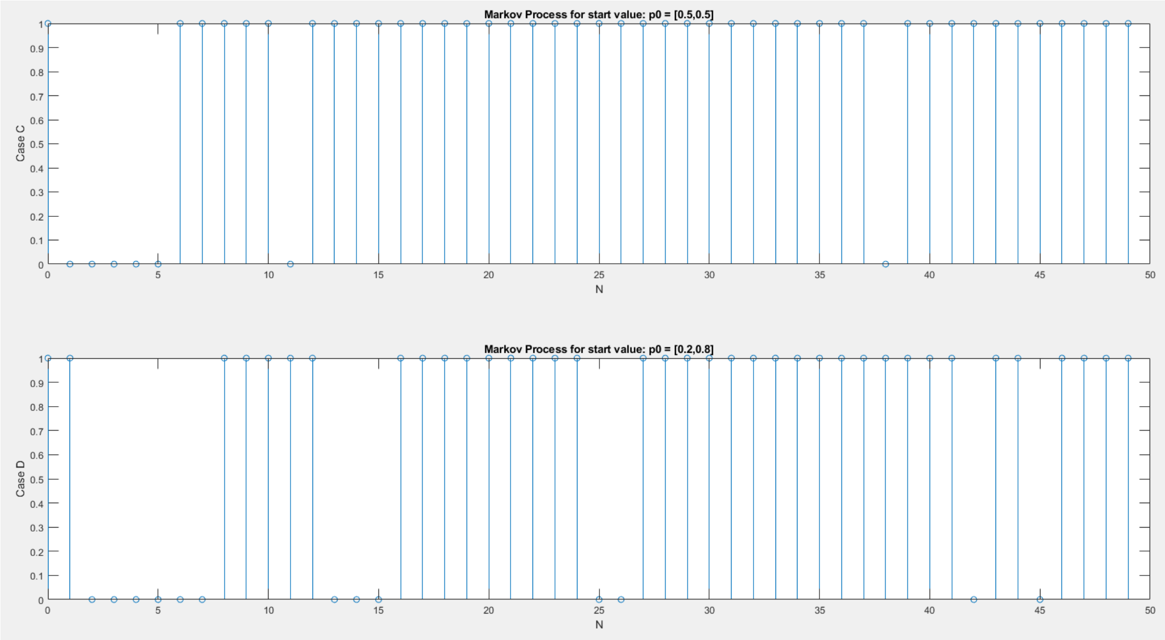
<u>סעיף 2:</u>



<u>הסבר:</u> למעשה בFigure הנ"ל מתוארים תהליכים אקראיים ברנולי, אשר לכל מ"א מסויים יש הסתברות של 0.8 להיות 1 והסתברות של 0.2 להיות 0.
גם כאן ניתן לראות כי הגרפים שונים. הדבר נובע מהאקראיות של הגרלת 50 מספרים.

<u>המשך שאלה 5:</u>

<u>סעיף 3:</u>



<u>הסבר:</u> למעשה אנו רואים בFigure הנ"ל גרפים של שרשראות מרקוב.

שרשרת מרקוב הוא תהליך אקראי אשר הערך של כל מ"א (בזמן n מסויים) תלוי בערך המ"א הקודם.

במימוש שלנו הערך הראשון של השרשרת נקבע לפי ההסתברויות שבוקטור תנאי ההתחלה $p_0 = [1-p, p]$, כמ"א ברנולי.

מכאן ואילך ההסתברויות לקבלת ערך של כל מ"א הבא מותנות בערך הנוכחי, יחושבו גם כמ"א ברנולי במימוש שלנו. חשוב לציין כי התהליך אינו תהליך אקראי ברנולי, כיוון שכל מ"א הבא תלוי בערך של הקודם לו.

<u>המשך שאלה 5:</u>

<u>סעיף 4:</u>

להלן טבלת ההסתברויות לקבלת 1 לפי מספר ההגרלות בכל אחד מהתהליכים:

4×5 table

N	A	B	C	D
50	0.58	0.7	0.72	0.94
100	0.55	0.81	0.8	0.91
1000	0.483	0.791	0.794	0.822
10000	0.5008	0.801	0.8001	0.8023

<u>הסבר:</u>

תהליך אקראי A – תהליך אקראי ברנולי מסעיף א'.

תהליך אקראי B – תהליך אקראי ברנולי מסעיף ב'.

תהליך אקראי C – שרשרת מרקוב עם תנאי התחלה $p_0 = [0.5, 0.5]$.

תהליך אקראי D – שרשרת מרקוב עם תנאי התחלה $p_0 = [0.2, 0.8]$.

המסקנה העיקרית מטבלה זו, שעבור מספר קטן של הגרלות (לדוגמא N=50) האקראיות משחקת תפקיד חזק יותר מאשר במספר גבוה של הגרלות (לדוגמא N=10,000).

לכן, עבור מקרים A,B ככל שמספר ההגרלות גדול יותר הוא מתייצב לערכי האמת לקבלת 1.

במקרה של A ההסתברות מתכנסת בקירוב לערך האמיתי לקבלת 1 – 0.5.

ובמקרה של B ההסתברות מתכנסת בקירוב לערך האמיתי לקבלת 1 – 0.8.

ניתן לראות בנוסף כי C,D מתכנסים לאותה ההסתברות לקבלת 1, ללא תלות בתנאי ההתחלה עבור מספר מספיק גדול של הגרלות.

עבור $n \to \infty$ ההסתברות לקבלת 1 תתייצב לערך של 0.8.

הצבנו בmatlab n=10000 בנוסחה הבאה:

$$p[n] = p[0] * P^n$$

כאשר P זו המטריצה של ההסתברויות:

$$P = \begin{bmatrix} 0.6 & 0.4 \\ 0.1 & 0.9 \end{bmatrix}$$

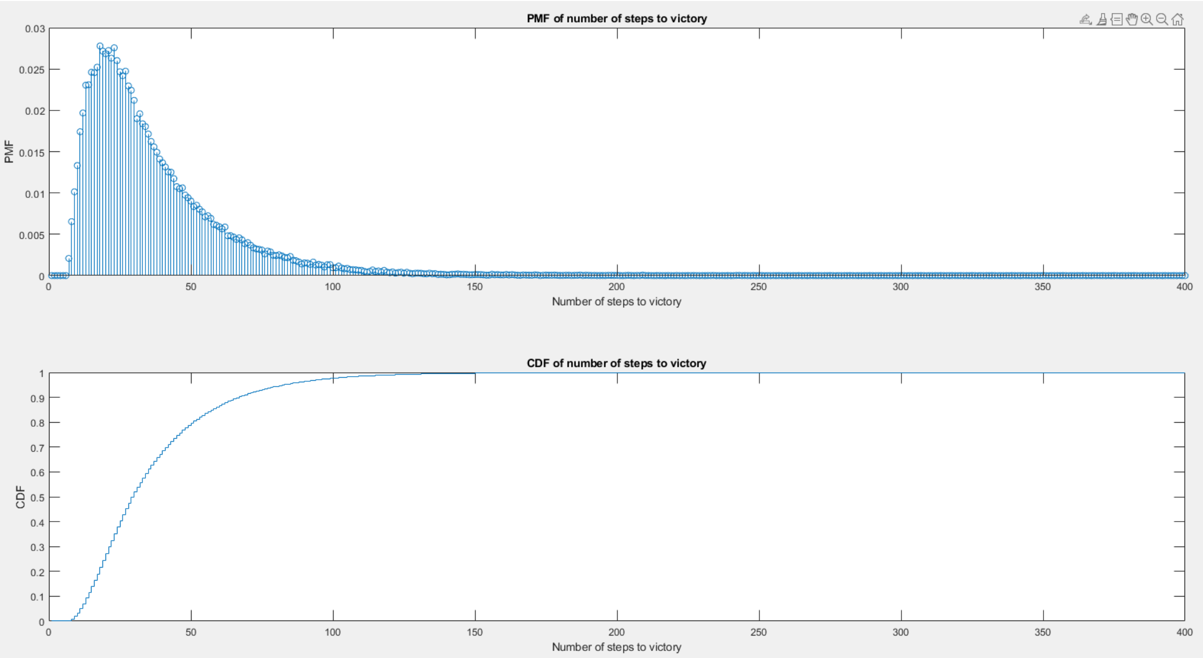
```
markov_max_C =

    0.2000    0.8000


markov_max_D =

    0.2000    0.8000
```

ניתן לראות כי ההסתברות לקבלת 1 עבור מקרים C,D כאשר $n \to \infty$ זהה לתוצאות בסימולציה מסעיף 4.

שאלה 6: קבצי מטלב נמצאים בתיקייה Q6

קישור לסרטון הסבר:
https://drive.google.com/uc?export=download&id=12LJ9xStSnC-yJVA3yGhCKnOJindNbOvI

הגרפים שמתארים את פונקציות ה-PMF,CDF מוצגים להלן:



מספר צעדים ממוצע לנצחון: 36.1259.

הסבר: ניתן לראות שבPMF ההסתברות הגדולה ביותר לנצחון היא סביב אזור ה19 צעדים (גרף עליון). חשוב לציין שהPMF אינה פונקציית הסתברות של תהליך אקראי גאוסי, למרות הדמיון בצורת הפעמון הגאוסי. ההבדל הוא שהממוצע הוא פחות או יותר סביב 36 צעדים מתחילת המשחק לניצחון. כלומר התוחלת לא בנקודת המקסימום של הפונקציה.

ראשית, פונקציית הCDF של תהליך אקראי בדיד (כמו אצלנו) היא פונקציית מדרגות (לא רציפה, הגרף התחתון). הפונקציה מגיעה להסתברות שקרובה מאוד ל-1, כאשר נגיע למספר שגדול יותר מ-120, מכיוון שהפונקציה עונה על השאלה: מה ההסתברות שלי לנצח אחרי מספר n>120.

שרשרת מרקוב למעשה היא תהליך אקראי אשר התפלגות של מ"א בתהליך תלויה בערך של המ"א הקודם לו. כלומר, ההתפלגות של X[n+1] תלויה בערך שיצא בX[n]. במשחק שלנו ההתפלגות של התא בהטלה הבאה תלויה במספר התא שאנו נמצאים בו. לדוגמא, אם אנחנו נמצאים בתא מספר 3, התא הבא שלנו יכול להיות בין 4-9 (בהנחה שאין סולמות ונחשים בתאים אלה). כלומר, התפלגות הערכים של התא הבא תלויה במספר התא שאנו נמצאים בו.