מבוא לעיבוד ספרתי של אותות ומידע

עבודת סיום

שמות המגישים: מתן שמש , אריאל כהן

כתב התחייבות לעבודה עצמאית:

חתימות: Ariel Cohen Matan Shemesh

מספרי תעודות זהות: 322624750 318449501

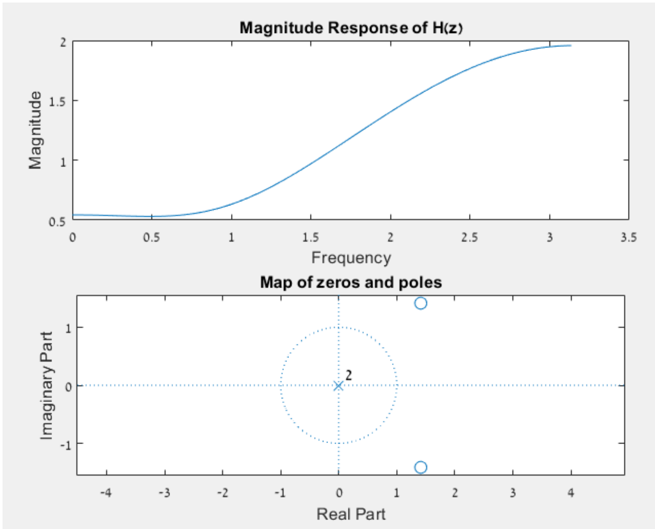
סרטוני הפרויקט:

חלק א':	חלק ב':
שאלה 1 שאלה 2 שאלה 3 שאלה 4 שאלה 5	פרק א' פרק ב' חלק 1 פרק ב' חלק 2 פרק ג'

חלק א'

שאלה 1 ([סרטון הסבר לשאלה](#))

א.



כפי שניתן לראות בגרף, מדובר במסנן HPF, תדרים נמוכים לא עוברים וגבוהים כן.

המסנן הוא FIR. כפי שלמדנו בכיתה, כאשר p=0 (מקדמי המכנה), נקבל פונקציית תמסורת ללא המכנה אשר במקרה שלנו היא תהיה מהצורה: $H(z)=\frac{1}{4}-\frac{\sqrt{2}}{2}z^{-1}+z^{-2}=b_0+b_1z^{-1}+b_2z^{-2}$ (כאשר $b_0=\frac{1}{4}, b_1=-\frac{\sqrt{2}}{2}, b_2=1$)

משך התגובה להלם שלה הוא סופי: $h[n]=\begin{cases} b_n\,, & 0\le n\le 2 \\ 0 & ,\ else\end{cases}$

א.

$$H(z)=\frac{1}{4}\frac{z^2-2\sqrt{2}z+4}{z^2}=\frac{1}{4}\frac{\cancel{z^2}(1-2\sqrt{2}z^{-1}+4z^{-2})}{\cancel{z^2}}=\frac{1}{4}(1-2\sqrt{2}z^{-1}+4z^{-2})=\frac{1-2\sqrt{2}z^{-1}+4z^{-2}}{4}$$

$a=[4,0,0]$ $b=[1,-2\sqrt{2},4]$. אין אפסים במעגל היחידה ולכן ניתן לז"ג נאמנות: $H(z)=H_{mp}(z)H_{Ap}(z)$

$$H(z)=\frac{1}{4}\frac{z^{-2}}{(z^{-1}-(\sqrt{2}+\sqrt{2}i))(z^{-1}-(\sqrt{2}-\sqrt{2}i))}$$ (הופכת השבר המקורי והוצאת שורשים) נחשב תחילה לביטוי המקורי ונמשיך את הפיתוח מעט.

$$=\frac{1}{4}\frac{z^{-2}}{1-2\sqrt{2}z^{-1}+4z^{-2}}=\frac{1}{4}\cdot\frac{\cancel{z^{-2}}}{\cancel{z^{-2}}(z^2+2\sqrt{2}z+4)}=\frac{z^{-2}-2\sqrt{2}z^{-1}+4}{4}$$

$$H(z)_{mp}=\frac{z^{-2}-2\sqrt{2}z^{-1}+4}{4}$$

$$=\frac{1}{4}\frac{z^{-2}}{(z^{-1}-(\sqrt{2}+\sqrt{2}i))(z^{-1}-(\sqrt{2}-\sqrt{2}i))}\cdot\frac{(1-(\sqrt{2}+\sqrt{2}i)z^{-1})(1-(\sqrt{2}-\sqrt{2}i)z^{-1})}{(1-(\sqrt{2}+\sqrt{2}i)z^{-1})(1-(\sqrt{2}-\sqrt{2}i)z^{-1})}$$

$$H(z)_{Ap}=\frac{(1-(\sqrt{2}+\sqrt{2}i)z^{-1})(1-(\sqrt{2}-\sqrt{2}i)z^{-1})}{(1-(\sqrt{2}+\sqrt{2}i)z^{-1})(1-(\sqrt{2}-\sqrt{2}i)z^{-1})}=\frac{1-2\sqrt{2}z^{-1}+4z^{-2}}{4-2\sqrt{2}z^{-1}+z^{-2}}$$

$$H(z)=H(z)_{mp}\cdot H(z)_{Ap}=\frac{z^{-2}-2\sqrt{2}z^{-1}+4}{4}\cdot\frac{1-2\sqrt{2}z^{-1}+4z^{-2}}{4-2\sqrt{2}z^{-1}+z^{-2}}$$

שאלה 2 (סרטון הסבר לשאלה)

h[n] הוא מסנן LPF בעל תדר קיטעון של $\pi/2$

ג. כאשר נפעיל את x1,x2 נוכל לשמוע את האותות המקוריים שקיבלנו בשאלה.

Z מהווה חיבור של x1 יחד עם x2 (אשר בכל פעם יוכפל במקדם חיובי או שלילי לפי אינדקס n, אשר יוצרת עיוות של האות המקורי, התבצע שינוי סימן לכל תדר אי זוגי. כמו כן, כתוצאה מההכפלה יש הזזה בתדר), כאשר ניתן להבחין כי האות של x1 הוא הדומיננטי, לפי השרטוטים באזור האמצעי עוצמתו של x1 חזקה יותר וכמו כן בתחילת האות שומעים רק את x1, ולאחר מכן שומעים בצורה מעומעמת את x2, ואכן תוצאה זו צפויה בגלל ההכפלה במקדם מינוס אחד.

נוכל להבחין כי האותות לאחר ההכפלות במסננים עברו הזזה ציקלית , דבר הגורם לתדרים הגבוהים להישמע ולתדרים הנמוכים להיות מאוד מעומעמים.
במגניטודה של y1 לאורך כל האות יש עוצמות גדולות ורחבות יותר מאשר של y2
בפונקציה של y1 האות הדומיננטי הוא x1, ויש דמיון רב לגרף של z.

ההכפלה במקדם חיובי או שלילי לפי אינדקס n בפונקציה y2 מפצה על ההכפלה ההתחלתית שבוצעה בz, לאחר שבוצעה הזזה בתדר פעם אחת יש הזזה נוספת. כעת האות הדומיננטי הוא x2, ואכן יש דמיון לגרף של x2, השמע של x1 יהיה ברקע בעוצמה חלשה מאוד. כל זה תוצאה של הנחתת תדרים גבוהים של X1

ו. קיבלנו תוצאות דומות במערכת 1 ובמערכת 2. במערכת 2, בעקבות תגובת ההלם שהעברנו את האותות בהתחלה הצלילים נשמעים מעט יותר טוב בz,y1,y2 לעומת מערכת 1. עם זאת, מעבר במסנן לפני חיבור 2 האותות. ניתן לראות כי y2 דומה בשני המערכות בגלל שהאות x2 יותר דומיננטי באות שלו, ולעומת זאת בy1 קיים הבדל בין 2 המערכות כי כעת האות הדומיננטי הוא x1.

ההכפלה בפעם השניה במסנן מנחיתה שוב תדרים גבוהים.
פעולת הקונבולוציה של x1 ושל x2 עם המסנן מנחיתה תדרים גבוהים.
שוב, גם בסעיף זה ההכפלה במינוס אחד בחזקה יוצרת עיוות של האות לכל התדרים האי זוגיים. גם כאן השמעה של Z יהיה שילוב של צלילים כפי שראינו קודם.

ז. כפי שציינו בסעיפים הקודמים, כאשר מכפילים קודם כל במסנן ואז מחברים עם אותות אחרים, האות מועבר בצורה יותר טובה וגם הסאונד יותר טוב, גם הגרף נראה יותר ברור ויותר זהה לאחד מהאותות המקוריים, ולכן נעדיף להעביר קודם במסנן ורק אז ליצור את Z למשל.

קראנו מידע באינטרנט על הכפלת אותות שמייצגות שמע במסננים וראינו כי קיים קשר לנושא הסיבתיות שלמדנו בכיתה. מערכת 2 יכולה להיות טובה יותר מכיוון שהיא סיבתית (הכפלה במסנן בהתחלה), כלומר היא לא דורשת דגימות עתידיות כדי לחשב את הפלט. זה מקל על הטמעה במערכות בזמן אמת ומבטיח שלא יוכנס עיכוב או עיוות עקב המסנן.
מצד שני, הסינון הלא-סיבתי (הכפלה במסנן בסוף) המשמשת במערכת 1 יכולה להכניס עיכוב באותות המוצא ונכנס עיוות.

שאלה 3 (סרטון הסבר לשאלה)

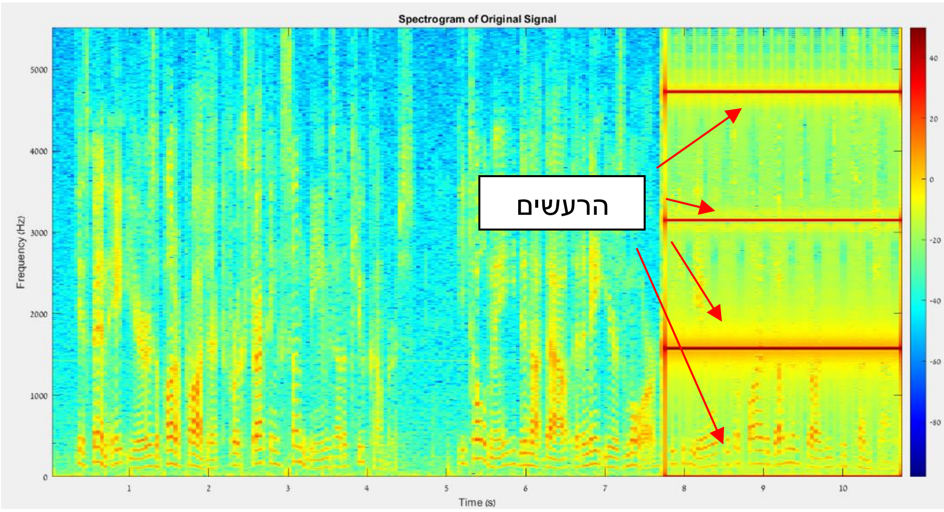
ב. מתוך הסתכלות על כל גרף וגרף, ניתן לראות בנקודות שבהן ערכו של ^x_c שווה לערכו של x_c השגיאה היא אפס, ובשאר המעברים בין הנקודות קיים שוני בין ^x_c ל x_c, ולכן יש ערך שונה מאפס עבור השגיאה. הסיבה שקיימת שגיאה רחבה לאורך כל הדרך היא שקשה להדמות ל"1' באמצעות פונקציית סינוס.
ככל שיש יותר איברים בסדרה N, כך גם אורך ציר הזמן t גדל.
במעבר מגרף לגרף, אפשר לראות דמיון בהתאם לטווחים, ז"א הגרף של N=10 מכיל בתוכו את הגרף של N=5 וכן הלאה ככל שN גדל. ככל שיש יותר איברים (N), כך השחזור יותר מדויק ודומה למקור.
השגיאה מתנהגת כפונקציה של T. אם נעביר קו באמצע, נראה כי פונקציית השגיאה היא זוגית (ניתן לראות התנהגות של מראה לשני הכיוונים).
כאשר נסתכל על הצורות של האותות, השגיאה שונה בכל מספר של איברים.
ככל ש-^x_c דומה יותר לx_c , כך השגיאה קטנה יותר ולהפך. ככל שיש יותר איברים N, כך השגיאה חזקה יותר או פחות.
^x_c הוא פונקציה שמאוד דומה לx_c ובעזרת מספיק איברים נוכל לומר שהם זהים ושזה שחזור מספיק קרוב.

ג. מתוך הגרף נוכל להסיק שככל שערכו של N גדול יותר, כך גודל השגיאה קטנה יותר. ניתן לשים לב שבעבור מספר N זוגי, הוא יהיה גדול יותר משכניו האי זוגיים, וזאת מכיוון שמדובר בפונקציית סינוס. גרף השגיאה קופץ בצורה זו מכיוון מדובר בפונקציית סינוס.
בגלל שאנו לוקחים את הערך המקסימלי מהשליש האמצעי, ונוצר בכל פעם כפולות אחרות בתוך הסינוס, אז פעם אחת הוא יפלוט ערך גבוה יותר או נמוך יותר (ביחס לכפולות של פאי).
נראה כי ככל שיש יותר איברים השגיאה קטנה לאט לאט.
ערך הN משנה את הפאזה באותו שליש מדובר, ולכן האמפליטודה עולה ויורדת בצורה קיצונית.

ד. נזכור כי המטרה של הDAC הוא להמיר להיות כמה שיותר דומה למקור. בסעיף זה בניגוד לנתונים הקודמים, כבר מההתחלה קיים חפיפה טובה יותר בין המקור להמרה, ואז מכיוון ש- x_c הוא קוסינוס, יותר קל להדמות איתו לסינוס.
ככל שמתקדמים על הציר, ^x_c מתחיל להיות דומה יותר ויותר לx_c, ובאמת כאשר יש N גבוה, השחזור נראה כמעט שחזור מושלם, והשגיאה היא מזערית.
בגרף השגיאה המקסימלית, אפשר לראות שערכי השגיאות קטנים יותר מהנתונים הקודמים, מהסיבות שפרטנו לעיל וכמו כן ככל שמתקדמים השגיאה קטנה וגם השיפוע מכמות איברים N אחת לאחרת. בנוסף, קיימת 'צניחה' גדולה מהירה יותר בגרף זה לעומת הנתונים הקודמים כי מדובר בקוסינוס וסינוס.

שאלה 4 (סרטון הסבר לשאלה)

א. ספקטרוגרמה היא צורת הצגה גרפית של תדרים של אות לאורך השתנתו בזמן.
לאחר שטענו את האות ואת קצב הדגימה מהאות שבקובץ, השתמשנו בפונקציית הספקטרוגרמה כדי לחשב את הספקטרוגרמה של האות. לפי מידע שקראנו באינטרנט, היה צריך לקבוע מאפיינים נוספים לאות אשר יאפשרו לפונקציית הספקטרוגרמה לקרוא את המידע ולעבד אותו (window,noverlap,nfft) , ככל שהנתונים שנזין עבור ערכים אלה יהיו גדולים יותר, כך הפסים של הרעשים יהיו דקים יותר ויהיה ניתן לקרוא אותם ביתר קלות.
פונקציית הספקטרוגרמה מחזירה 3 ערכים: S-מהווה המרת פורייה. ,F-התדרים המחזוריים אשר שווה למספר השורות של S, T-קטעי זמן התואמים לנקודת אמצע של כל קטע.
לאחר הקריאה של האות והתבוננות על העוצמות בכל רגע ורגע, ציירנו את הספקטוגרמה באמצעות imagesc.
התדרים עם האנרגיה הגבוהה באות הם האותות הסינוסידיאלים, תדרים אלו הודגשו בגרף ומתוך התבוננות מדויקת בערכיהם, יכלנו להסיק מהם אותם תדרים (בקוד קראנו להם freqs).
לאחר מכן, ביצענו המרה של כל תדר ותדר מתדר בהרץ לתדר בזמן בדיד.



ב. בהינתן h, באמצעות חישוב מתמטי, פיתחנו נוסחה למציאת A. בסעיף הקודם **תוך התבוננות בגרף מצאנו את תדרי הרעש** וכעת כל שנותר לעשות הוא להציב בנוסחא שפיתחנו את ערכי התדרים ולמצוא כל ערך A באמצעות ריצה בלולאה על התדרים.

$$h=[1,A,1]$$

$$h(n)=\delta(n)+A\delta(n-1)+\delta(n-2)$$

$$H(w)=1+Ae^{-jw}+e^{-2jw}=0$$

$$A=-\frac{1+e^{-2jw}}{e^{-jw}}=-\frac{\cancel{e^{-jw}}\cdot e^{jw}+\cancel{e^{-jw}}\cdot e^{-jw}}{\cancel{e^{-jw}}}=-e^{jw}-e^{-jw}$$

$$A=-2\cos(w)=-2\cos\left(\frac{2\pi f}{f_s}\right)$$

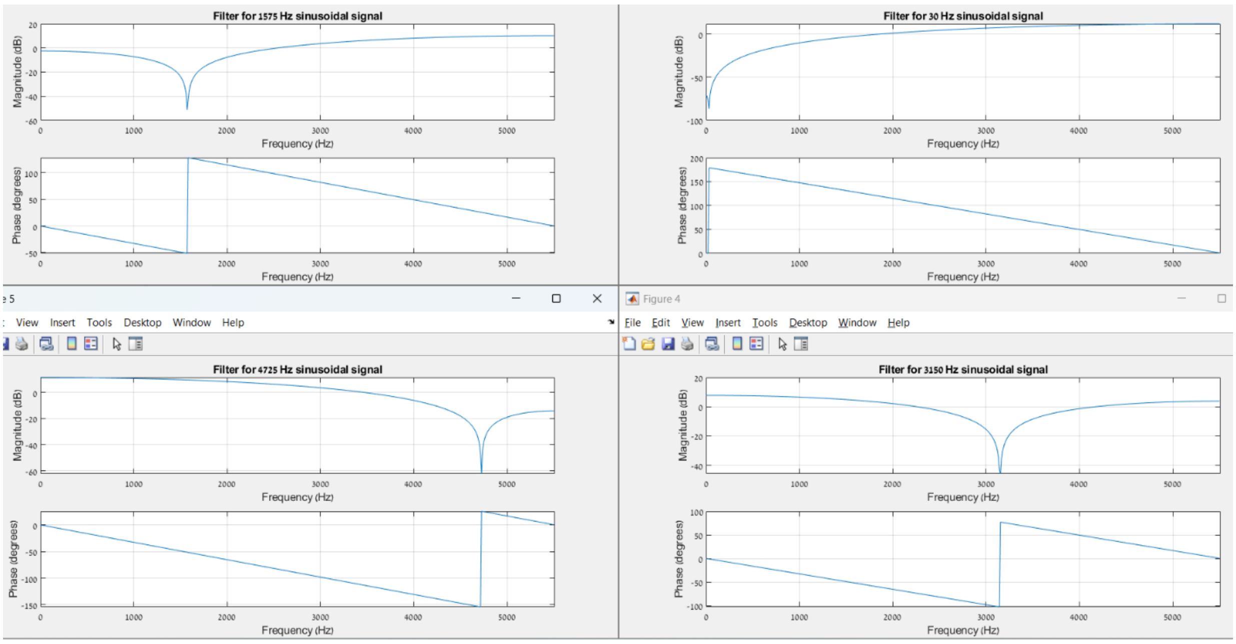
$$w=\frac{2\pi f}{f_s}$$

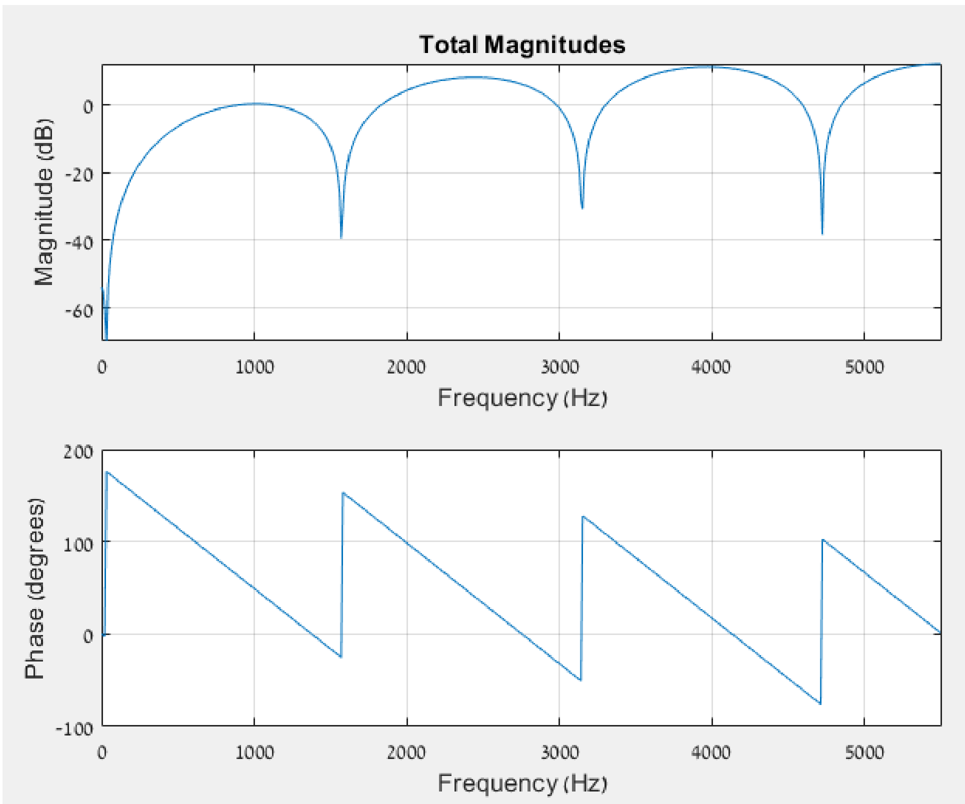
תדר דגימה

ג. להלן הטבלה המבוקשת שאת ערכיה חישבנו בקוד:

HZ	w	A
30	0.3682	-1.999
1575	19.3282	-1.246
3150	38.6563	0.445
4725	57.9845	1.801

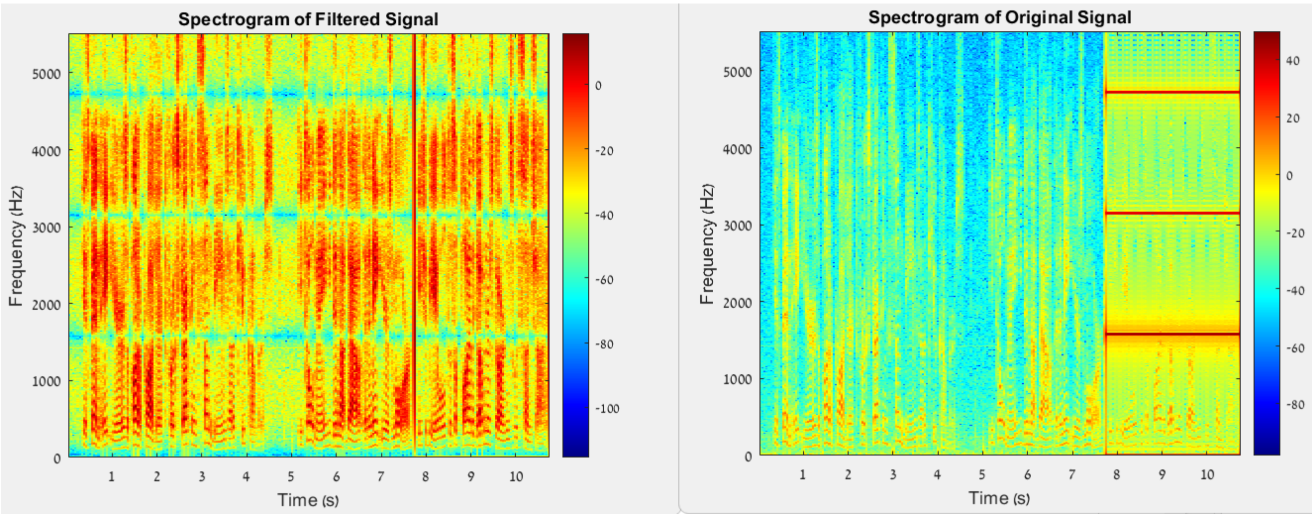
ד+ה.





ו. לפני הסינון יש בידינו אות שמכיל דיבור, ובשלב מסוים נכנסים רעשים שדורסים את נוכחותו של אות הדיבור. לאחר הסינון, קיבלו אות שנשמע מעט יותר מעומעם, אך לא מכיל את הרעשים, ניתן להבחין בנקודה שבה היה הרעש בעקבות רעש קטן שנשמע ברקע. קיבלנו אות נקי (יחסית) ללא רעשים, שבתהליך כאן בשאלה הסרנו את הרעשים.

ז.



בשרטוט לפני ניקוי הרעשים, ניתן לראות את האותות הסינוסואידלים שמהווים רעשים בצד ימין של הגרף (בסעיפים הקודמים כך מצאנו את הרעשים תוך התבוננות בהם). עשינו שימוש בפונקציית imagesc שמאפשרת לצייר גרף ספקטוגרמה, וכמו כן הוספנו את המאפיין colormap(jet) אשר התדרים הגבוהים ביותר מסומנים בצבע אדום ובולט יותר, והתדרים הנמוכים בצבע ירוק וכחול. לפני הסינון, אפשר לראות שהאדומים הם האותות הסינוסואידלים, ולאחר הסינון כמעט כל הגרף צבוע באדום מהסיבה שכעת כל התדרים בערך באותה עוצמה.

שאלה 5 ([סרטון הסבר לשאלה])

ד. ניתן לראות כי השרטוטים של האות המקורי ושל השגיאה דומים, גם האות וגם ההתמרה, זאת בגלל שהחסרנו מהאות המקורי רק 10 אחוז ולכן נשמור רוב מוחלט של התדרים והעוצמות.

ההבדל בין האות המקורי לאות המשוחזר מתרחש בגלל אובדן המידע במהלך תהליך הדגימה מכיוון שחילקנו את האות המקורי לחלקים .האות המקורי מחולק ל-20 חלקים שווים, ולאחר מכן ה-DFT נלקח עבור כל חלק. מתבצע חישוב DFT לכל חלק בנפרד. בחישוב ה-DFT, נחשבים רק 10% מהערכים הראשונים. כתוצאה מכך יש אובדן מידע כי לא לקחנו את כל האות המקורי בתהליך. בסעיף ג', חיברנו כל חלק שביצענו לו DFT בנפרד לחלק אחד על מנת שנוכל לצייר אותו. השחזור אינו מדויק בגלל אובדן מידע שהתרחש בשמירת 10% מהערכים הראשונים. לכן, האות המשוחזר מכיל עיוות בהשוואה לאות המקורי.

האות המשוחזר נשמע עמום יותר בגלל איבוד המידע, וגם הגרף שלו נראה מעוות יותר ונראה כמו שקיבל 'צביטה' והוא פחות דחוס. כתוצאה מכך אנו שומעים כי הסאונד נשמע מעט אחרת וזה בעקבות כל התהליך שהוא בהתמרה ובהתמרה החוזרת, ישנם מקומות שהיו בהם ערכי אפסים כי נלקח רק 10 אחוז מהמקור והם לא ישפיעו בשמע ובתצוגה של הגרף בהתמרה החוזרת.

ה. אות השגיאה e[n] מתקבל על ידי הפחתת האות המשוחזר מהאות המקורי. אות השגיאה מייצג את ההבדל בין שני האותות, והוא מכיל את מידע התדר הגבוה שאבד ואת העיוות, ז"א שהוא יכיל 90% מהאות המקורי בחישוב מתמטי ישיר. גרף השגיאה זהה לגרף המקורי וקיימים בו אזורים שבוצעה בו הנחתה של האות בעקבות ההתמרות.

איור התמרת הפוריה של השגיאה דומה לזה של האות המקורי, כאשר אותם 10 אחוז שנלקחו כעת מיוצגים בתור אפסים ולכן אחרי ההתמרה הם לא יהיו אפסים גם כן ולא יוצגו עם ערכים.
ניתן לראות בגרף שהוא דומה מאוד לאות המקורי וגם ההתמרה שלו וזאת מכיוון שהוא אותו אות רק עם איבוד של 10%. מכאן שגם הסאונד של אות השגיאה ישמע דומה מאוד לאות המקורי עם מעט עיוות.

חלק ב' – אימון מסווגים בינאריים

פרק א: אימון מסווג בינארי מסוג KNN לסיווג אותות סונאר (סרטון הסבר לפרק)

שאלה 6

ב. נדאג לכך שבסדרת האימון ובסדרת המבחן יהיו גם דוגמאות של y=0 וגם דוגמאות של y=1, על מנת שהאלגוריתם ידע להתמודד עם סיווג שונה. נפצל את המידע ביחס של 1:4. הסיבה לבחירה בפיצול זה היא להבטיח שהמודל יאומן על כמות גדולה מספיק של נתונים ובכך יכיר את הדפוסים והקשרים הבסיסיים בנתונים. במילים אחרות, אנחנו רוצים לתת לו מספיק דוגמאות של נתונים לבדיקה מעמיקה והערכת ביצועי המודל. בנוסף, פיצול זה עוזר לנו מבחינה הסתברותית למנוע חזרה של אותן דוגמאות בסדרת המבחן.
בעבור סדרת האימון, יהיו לו 81 דוגמאות של אפסים ו85 דוגמאות של אחדים.
בעבור סדרת המבחן, יהיו לו 16 דוגמאות של אפסים ו26 דוגמאות של אחדים.
N_train=166, N_test=42

שאלה 7

בעבור K=3,
כאשר נכניס לאלגוריתם את ערכי X ו-Y של סדרת האימון ואת X של המבחן, נצליח להגיע לדיוק של 88.1% .
כאשר נכניס לאלגוריתם את ערכי X ו-Y של סדרת המבחן ואת X של האימון, נצליח להגיע לדיוק של 61.45% .
אנו מבינים שיש להפעיל את האלגוריתם של הKNN עם סדרת האימון ולבדוק את סדרת המבחן, התהליך ההפוך לא נותן אחוזי דיוק גבוהים כי זה לא המטרה של הKNN לפני הבנתינו.

שאלה 8

נתחיל מהנחת בסיס שאין לרוץ על ערכי K זוגיים, מכיוון שיכול להיווצר מצב של שיוויון בין מספר האפסים למספר האחדים והאלגוריתם לא ידע כיצד לסווג.
הרצנו בלולאה את כל ערכי K האפשריים החל מ1 ועד הגודל של Ytest, ומצאנו כי בעבור k=1 או k=3 נקבל את הדיוק המירבי של 88.1%.

פרק ב: אימון מסווג מסוג רגרסיה לוגיסטית (סרטוני הסבר לפרק: חלק 1 , חלק 2)

שאלה 9

ניצור בשלב הראשוני וקטור שורה W באורך הסדרה X, ונגריל לתוכו מספרים אקראיים בין 0 ל-1.
לאחר מכן, נקבע מקדם אלפא ואת מספר האיטרציות הרצויות, ועל כל איטרציה כמספר האיטרציות – נקבל וקטור חדש של W שמספר השורות שלו הוא כמספר האיטרציות ומספר העמודות כשל X.
בלולאה כמספר האיטרציות, נפעיל את האלגוריתם של gradient descent, שכל חלק שולח אותו לפונקציה אחרת שתורמת לחישוב (מפורט בקוד).

שאלה 10

א. לאחר שבסעיף הקודם יצרנו האלגוריתם למציאת סדרת W, השתמשנו באלגוריתם בסעיף זה כאשר ביצענו 400 איטרציות וקבענו את מקדם אלפא להיות 0.01 (קביעת מספר האיטרציות ומקדם אלפא היה ניסוי וטעיה).
לאחר מכן, רצנו כמספר האיטרציות שנקבע, בחרנו מקדם trashold אידאלי של 0.5 ובכל שלב השתמשנו

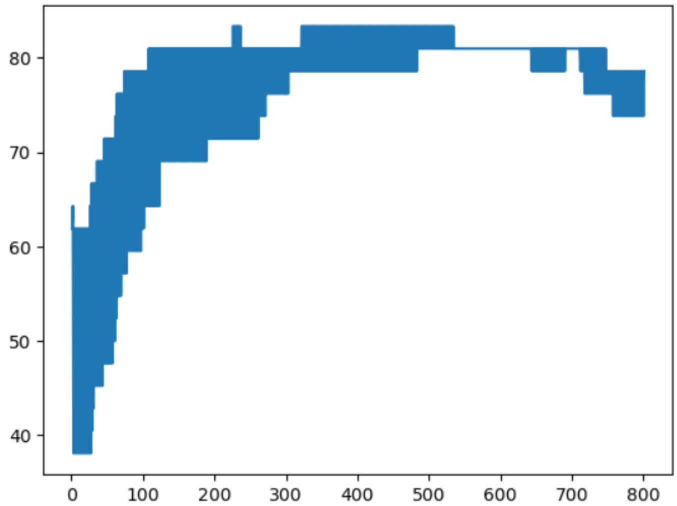
בפונקציות probablisticLogRegClassifier ו-FinalClassification , על מנת לחשב את הדיוק, ובמקביל הדפסנו את גרף הדיוק הסיווג.
לשלב הבא, ניגשנו לטפל ב Cross Entropy באמצעות הפונקציה המיועדת ועדכנו את גרף פונקציית הקנס בכל מעבר של איטרציה של W.
ניתן לראות שכלל שביצענו יותר איטרציות, כך ערך הCE יורד. לבסוף ציירנו את גרף פונקציית הקנס.

ב. בהמשך לסעיף א', השתמשנו באותה W שמצאנו בעזרת סדרת האימון בשביל לשמר על אחידות המודל, הצבנו את אותם מספר איטרציות, מקדם אלפא וtrasholdו וקיבלנו את התוצאות המבוקשות.

שאלה 11

כפי שלמדנו ונסינו במעבדה, גם כאן אנו יודעים שעבור ערך T קטן מידי לא יהיה דיוק מספיק למסווג, ורק לאחר מספר גדול מספיק נראה אחוזי דיוק גבוהים. כפי שאפשר לראות בגרפים שאנו מציגים, בעבור T גדול מידי לא יהיה שינוי באחוזי הדיוק וזה נחשב לבזבוז משאבים. מצאנו שבעבור פרמטרים של אלפא של 0.01, T=400, th=0.5 אידאלי ניתן להגיע לדיוק הטוב ביותר בעבור סדרת האימון, וגם בעבור סדרת המבחן.

לעומת זאת, אם נבחר את אותם נתונים וT=800,זה יכול לשפר את סדרת האימון באחוזים בודדים אך בעבור סדרת המבחן נקבל את התוצאה הטובה ביותר לאחר כ400 ריצות, ומשם חלה ירידה משמעותית בדיוק כפי שניתן לראות למשל בגרף זה:



בפרויקט שלנו, יצרנו פונקציה שמחזירה את W, כך שבכל פעם שמפעילים את הפונקציה היא מגרילה מספרים רנדומליים בין 0 ל-1, וכך באמת אנו בוחנים בשאלה זו את השפעת ערכי W ההתחלתיים וכתוצאה מתהליך זה ראינו כי קיימת השפעה של אחוזים בודדים על אחוזי הדיוק.

לW ההתחלתי ((w(0)) אכן יש השפעה על מספר הצעדים הנדרשים T לקבלת אחוזי הדיוק הגבוהים ביותר.
כפי שתיארנו למעלה, נציג כעת דוגמא עם הרצה מקסימלית של T=800, וכמו כן a=0.01.
לחלק זה יצרנו פונקציה שנקראת Impact_Test, אשר עושה הרצות שונות עם ערכי T שונים וערכי אלפא שונים.

מכיוון שאצלנו W ההתחלתי הוא רנדומלי, הרי שברור שתהיה לו השפעה על ערכי W הסופיים.

בשביל לבדוק את השפעת W ההתחלתי ואלפא על מספר הצעדים הנדרשים T, ביצענו את המבחן הבא:

```
def Impact_Test (a):
    Xtrain,Xtest,Ytrain,Ytest= _Myinit()
    T=[1,10,100,400,800]
    for t in T:
        print("\n\033[1;34mIn this stage , a= "+str(a)+" , T= "+str(t)+"\033[0m")
        W_coff = calc_w_coff(Xtrain,Ytrain,a,t)
        P_arr = Classification_accuracy(Xtrain,Ytrain,W_coff,0.5)
        print("\n\033[1;33mThe first value of W[0] in this stage: \n" +str(W_coff[0])+"\033[0m")
        print("\n\033[1;32mThe final value of W in this stage: \n" +str(W_coff[len(W_coff)-1])+"\033[0m")
```

```
a=[0.01,0.05,0.1,0.3,0.5,1]
for i in range(len(a)):
    Impact_Test(a[i])
```

בכל לולאה, ביצענו הרצה עם ערכי אלפא שונים, שבעבור כל אלפא רצה לולאה עם ערכי T שונים,
במהלך כל ריצה הדפסנו את אחוזי הדיוק בכל שלב T פנימי, והדפסנו את מערך W הסופי.

בקובץ txt כדוגמא כאן (או כאן) אספנו את הדיוק המקסימלי בעבור כל אלפא ועבור כל T.

בעבור ערכי אלפא קטנים וT קטנים, כבר בשלב הראשוני של אותו T נקבל את הדיוק המקסימלי האפשרי ואין צורך בהמשך השלבים (למשל בזמן a=0.01, T=1 or T=10)
בעבור ערכי אלפא קטנים וT גדולים אכן יש יעילות במספר ההרצות, שכן שכלל שמגיעים להרצות האחרונות של כל T, רואים את הדיוק המקסימלי.
לכל מקרה, כאשר נציב T גדול מידי (למשל 1000) נגלה שאין בכך יעילות ואחוזי הדיוק יורדים או לא משתנים.

עבור ערכי אלפא "בינוניים" (למשל a=0.1), עדיין נקבל דיוק גבוהה יותר לאחר מספר סביר של הרצות T, אבל מגיעים לדיוק מקסימלי לרמה זו לפני שמגיעים לקצה התחום.

עבור ערכי אלפא גבוהים, נראה התנהגות דומה לזו של ערכי אלפא "בינוניים".
וגם אפשר לראות שאין יציבות במהלך האיטרציות (כמו שניתן לראות בקובץ הtxt לדוגמא), ז"א שיהיו קפיצות משמעותיות בכל שלב ושלב ולא יהיה ניתן להבין את הדיוק כבר בשלב זה.

באופן כללי, ניתן לומר שרק עבור ערכי אלפא קטנים מספיק (למשל 0.01) נגיע לאחוזי דיוק גבוהים באזור של כ-85%.

כמו כן, השוונו בין ערכי W ההתחלתיים לבין ערכי W הסופיים בעבור אלפא קטנים וגדולים,

כאן ניתן לראות כאשר a=0.01,T=500, ועם וקטור W התחלתי רנדומלי:

```
The firsst value of W[0] in this stage:
[0.71827076 0.2553748  0.12719261 0.84747172 0.15256991 0.15929578
 0.31085117 0.60184208 0.76348749 0.90597032 0.93778033 0.35978965
 0.20137506 0.79517205 0.47298306 0.95150429 0.07601033 0.32099062
 0.57439778 0.30385406 0.52520655 0.02914903 0.79221515 0.00644092
 0.33088175 0.70666518 0.40275586 0.74983255 0.58279929 0.48183338
 0.48711914 0.36860518 0.30378016 0.30474228 0.01591709 0.83253991
 0.33548334 0.99641347 0.95586313 0.43214922 0.96223778 0.40882594
 0.42229553 0.95132215 0.47985261 0.91984002 0.1661008  0.74845842
 0.08431792 0.20450739 0.98568246 0.24590567 0.73388044 0.33661245
 0.01359212 0.96433476 0.91767873 0.25943551 0.91893478 0.48031822
 0.81410607]

The final value of W in this stage:
[-3.85148176  1.14916792  0.62875641  0.7135522   1.70685111  1.77762947
  0.02383937 -1.03176567 -0.59458264  2.00930646  1.84207206  4.00412801
  2.92166561  0.54523113 -1.70864712 -0.62401778 -1.42494302 -0.63608523
 -0.4875678   0.98321747  0.99690815 -0.03462021  0.96536191  0.2409086
  1.87990766 -0.28991483 -1.2929315   0.20469941  0.36008559  0.15830147
  0.87724491 -3.71368726  1.26212253  1.16393188 -1.24883612 -0.4874102
 -2.41143579 -1.77252399  0.43134991  0.44846881 -2.57367866 -0.80693546
  0.82061449  3.31626631  2.93614204  3.929973    2.37080655  1.78882473
  1.65675029  1.54292308  0.92987092  0.65440879  1.07353515  0.49892476
  0.34212581  0.93948833  0.91898404  0.1487903   1.0926435   0.73083662
  0.90165455]
```

כאן ניתן לראות כאשר a=1,T=500, ועם וקטור W התחלתי רנדומלי:

```
The firsst value of W[0] in this stage:
[0.47565693 0.04286265 0.23778039 0.43846647 0.93066754 0.95566846
 0.38593855 0.6458314  0.91764797 0.70267884 0.61777237 0.5461845
 0.1012677  0.05364229 0.05819844 0.23240369 0.49270572 0.37614817
 0.86748462 0.22470335 0.11075964 0.47780667 0.86067681 0.80422859
 0.85067732 0.49502343 0.27894116 0.77957564 0.86467316 0.19267775
 0.20411024 0.37724039 0.06586644 0.76724504 0.57567574 0.60002676
 0.55555194 0.49782468 0.22172535 0.28892059 0.53616847 0.79030921
 0.30658591 0.60871136 0.8864592  0.25607268 0.72794857 0.98176093
 0.38062697 0.07816821 0.06638508 0.1382333  0.07891678 0.96201701
 0.8897896  0.35100293 0.11407711 0.23040979 0.74078137 0.16318442
 0.7557888 ]

The final value of W in this stage:
[-415.24765042  101.06922197   51.90287013   -9.98058738  172.43586777
  224.0306321    69.48301458 -101.95678971  -65.35709547  232.15620679
  246.76110513  483.80344028  338.42505224   71.23144473 -147.90800366
 -110.90987734 -163.91659815  -91.55079049  -39.63480973   96.89489468
  107.82429751   65.66693982   98.83214562  103.96423215  191.61537942
  -83.10516011  -64.46167785   23.91400334   84.33365515   71.58603139
  143.07124284 -395.45467514   93.76647961   57.32051925 -151.6806027
  -50.86172612 -182.83437281 -221.75535416   70.11619414   80.18377947
 -252.42882223  -28.38307524  118.73806133  346.07865042  335.36388013
  404.614127    294.32497414  141.18196574  184.65508248  151.25589973
    3.78384531   54.21225239   37.30721966   17.37384109   40.335545
   -2.14861357    1.98054727  -11.01957852   22.04548144   27.52769692
   11.97029656]
```

המסקנה מכאן היא שככל שאלפא יותר גדול, כך ערכי W הסופיים יהיו יותר גדולים.

בפרויקט שלנו, **חילקנו את המידע** בכל פעם באותה צורה של 1:4, אך יחד עם זאת ניתן בפונקציית הSplit להציב ערכים אחרים ב- random_state וזה ישנה לחלוקה אחרת והאחוזים יהיו כמעט אותו דבר בכל הרצה. גם כאשר נציב מספר אחר בrandom_state , נקבל את אותם אחוזי דיוק עם סטייה קטנה מאוד.

בשביל לענות על השאלה האם שגיאת האימון היוותה מדד אמין לשגיאת המבחן, לאחר חלוקת המידע כפי שמתואר לעיל, ביצענו הרצת בדיקת דיוק על סדרת האימון וקיבלנו (למשל) כ-80% דיוק, בעבור 166 דוגמאות, ז"א היה דיוק עבור 141 דוגמאות. לאחר מכן כאשר הרצנו בדיקה על סדרת המבחן קיבלנו (למשל) דיוק של כ-85%, עבור 42 דוגמאות, ז"א היה דיוק עבור 36 דוגמאות. אנו מבינים שככל שיש פחות דוגמאות לבדוק, אחוזי ההצלחה עולים, ואם קיבלנו אחוז הצלחה מסוים בסדרת האימון, נקבל אחוז דיוק דומה או מעט יותר גבוה בסדרת המבחן.

הוספנו בקוד גם ריצה על אחוזי דיוק של סדרת המבחן וניתן יהיה לראות התנהגות דומה.

כל ההיבטים לבחינת מהלך האימון צוינו בסעיפים הקודמים, בסעיף זה הנ"ל וגם בקוד עצמו.

פרק ב: השפעת נרמול המאפיינים על ביצועי האלגוריתמים (סרטון הסבר לפרק)

שאלה 12

בבחינת הKNN על סדרת האימון קיים שיפור – יש עליה מ88.1% דיוק לפני הנרמול לדיוק של 90.5% דיוק אחרי הנרמול.
כאן, כאשר K=3 נקבל את הדיוק הטוב ביותר (90.5% דיוק).

בבחינת דיוק המסווג,
מאחר ושאנחנו בנינו את המודל ואנחנו מכירים את התנהגותו, נוכל לקבוע את עניין היציבות, ז"א לומר איזה אינדיקציה נותנת לנו שהדיוק שקיבלנו במסווג הוא יציב ואמין.
לאחר הרצות רבות בחרו ששמירת יציבות תבוטא בכך שתשמור על אותו דיוק למשך כ-20 איטרציות.

בשלב זה, נקבע את ערכי w0 ההתחלתיים להיות אפסים (בחלק א' ערכיו אותחלו להיות מספר רנדומלי בין 0 ל1).
הסיבה לבחירת ערכי אפס הם בשביל חישוב gradient descent לאחר שלב הנרמול, אנחנו בעצם 'שוברים את הסימטריה' של ערכי W וכך גם מאפשרים לאלגוריתם להתחיל מ"ערכי אפס" כמו הממוצע ושונות.
לפיכך, לאחר תהליך הנרמול וחישוב gradient descent, המרחק בין הנקודות יהיה יותר טוב, ויהיה צורך בצעדים יותר קטנים (אלפא) מאשר הצעדים שעשינו לפני הנרמול ולפני הפיכת w0 להיות 0.

כעת, כאשר נשאר עם אותם ערכי th,a,T כמו לפני הנרמול אנחנו מגיעים לכמעט 100% דיוק בסדרת האימון, שזה כמובן לא תוצאה טובה. מצב זה יוצר אימון יתר וקיבוע לנתוני סדרת האימון, וזה עלול לפגוע בדיוק המסווג בעבור סדרת המבחן או בעבור דוגמאות חדשות נוספות. לכן אנו מבינים שצריך לשנות את הנתונים כדי שהמסווג לא יהיה מקובע ב100% על סדרת האימון (ז"א להכריח אותו להיות עם ביצועים מעט פחות גבוהים) ובכך נצליח לשפר את אחוזי הדיוק בעבור סדרת המבחן ודוגמאות נוספות.

לפיכך, קבענו a=0.001 במקום a=0.01 שהיה לפני הנרמול.
בעבור סדרת האימון, גילינו כי מתבצע בזבוז משאבים וישנן הרצות רבות מיותרות שלא משפרות כבר את אחוזי הדיוק של סדרת האימון (הוא נשאר על דיוק של 92.168% לאורך הרבה איטרציות נוספות) .ניתן לשנות את T=400 כפי שהיה לפני הנרמול לערך שלך T=250 ולהגיע לדיוק זהה מקסימלי של 92.168% ובכך בזבזנו פחות משאבים וזמן ריצה. תוצאה זו אשר מכילה אחוז דיוק גבוה יותר מלפני הנרמול מלמדת אותנו שאכן נרמול המידע לפני אימון המסווג משפרת באופן משמעותי את התנהגות המודל.

בסדרת האימון קיים שיפור, עלינו מדיוק של כ-80% לפני הנרמול לדיוק של כ-92% אחרי הנרמול.
בסדרת המבחן קיים גם כן שיפור, עלינו מדיוק של כ-80% לפני הנרמול לדיוק של כ-85% אחרי הנרמול.
כמו כן, בגרף שמייצר Classification_accuracy לפני הנרמול קיימים חלקים שמכילים 'זיגזגים' בגרף (כי יש קפיצות רבות בין ערכים גבוהים ונמוכים סמוכים) ולאחר הנרמול הגרף הוא 'חלק', יש התקדמות ולא מתעכב על אותם ערכים כל הזמן.

בבחינת Cross Entropy,
לפני הנרמול,
בסדרת האימון ערכיו נעים החל מערך כ-650 באיטרציה הראשונה לערך של כ-70 באיטרציה האחרונה.
בסדרת המבחן ערכיו נעים החל מערך כ-130 באיטרציה הראשונה לערך של כ-15 באיטרציה האחרונה.
לאחר הנרמול,
בסדרת האימון ערכיו נעים החל מערך כ-115 באיטרציה הראשונה לערך של כ-40 באיטרציה האחרונה.
בסדרת המבחן ערכיו נעים החל מערך כ-30 באיטרציה הראשונה לערך של כ-17 באיטרציה האחרונה.
המסקנה העולה מכך היא שלאחר הנרמול ההפרש בין מה שצפינו לקבל מהמסווג לעומת מה שקיבלנו בפועל כעת קטן יותר, הערכים לאחר הנרמול נעים בטווח יותר קטן. ניתן לראות שהגרפים יותר ברורים ולא קיים בהם חלקים

גסים של ערכים זהים כפי שהיה לפני הנרמול, הגרף יותר 'זורם'.

בבחינת גרפי הROC,
נראה שקיים שיפור באחוזי הדיוק של מטריצת הROC לאחר תהליך הנרמול, הרחבו על כך לעיל.
הגרפים כעת יותר חלקים, ניתן לראות בצורה ברורה יותר את התנהגות הROC בכל שלב של האיטרציות ואת קצב ההתקדמות של הדיוק.

לסיכום,
ניתן לראות כי קיים הבדל בתהליך האימון של המסווגים, נדרשנו לבצע התאמות בעבור W, בעבור מספר האיטרציות הנדרשות T, לשנות את מקדם אלפא ולהזין ערכים שונים בכדי להגיע לתוצאה רצויה.
תהליך זה תרם לנו רבות אודות תרומת נרמול הנתונים, הבנה מעמיקה יותר של המודל ואופן פעולתו וכיצד להימנע ממצב של קיבוע ומבזבוז משאבים.
כמו כן, קיים הבדל בדיוקים שהתקבלו, לאחר הנרמול אחוזי הדיוק עלו הן בעבור סדרת האימון והן ועבור סדרת המבחן, הראנו כאן בפירוט כיצד הדיוקים משתנים בכל פונקציה ופונקציה של הפרויקט.