ממ"ן 12 בחישוביות ביולוגית

# מגישה: סיניה נודל ת.ז 203191663

1. כדי למצוא ערכים אופטימליים x,y אשר מקטינים את הפונקציה g(x,y) אשתמש באלגוריתם גנטי בו:

א. כל כרומוזום מיוצג ע"י מחרוזת של 10 ביטים. 5 ביטים שמאליים עבור ערך x ו5 הביטים הימניים לערך. נתון שערכי x נעים בין 0-31 ולכן ניתן לייצג ע"י 5 ביטים את כל הערכים הללו. עבור ערכי y שנעים מ32 עד 64 נתייחס ל5 הביטים כאילו מיוצגים ע"י ביט 1 שערכו 1 נוסף משמאל. כך הייצוג ינוע בין 100000 (=32) לבין 111111 (=64).

ב. הכשירות של פרט (x,y) באוכלוסייה תחושב לפי . ערך C הוא קבוע חיובי והוא מייצג ערך מקסימלי שאני אעריך שהפונקצייה g(x,y) יכולה לייצר (לדוגמא: 1000C= ופרטים המייצרים ערכי פונקציה גבוהים מ1000 כלל לא ייחשבו כאופציות לפתרונות האפשריים). הקבוע מבטיח מצב שערכי פונקציה שליליים יקבלו ערכי כשירות גבוהים יותר מאשר ערכים חיוביים, ומבטיח שככל שערך פונקצייה חיובי גדול יותר – כך ערך כשירותו ייקטן.

ג. ההסתברות לבחירת כרומוזום היא המנה בין ערך הכשירות של הפרט לבין סכום ערכי הכשירויות של כלל הפריטים באוכלוסייה.

שיטת הסלקצייה שנבחרה היא הרולטה – בה לכל פרט סיכוי שווה להבחר כהורה בהתאם ליחס שבין כשירותו לסכום הכשירות של כל הפרטים באוכלוסייה. בשיטה זו יש סכנה לסחף גנטי והתכנסות מוקדמת כאשר האוכלוסיה מכילה פרטים בעלי פער ערכי כשירות גבוהים. שיטות נוספות יכולות להיות:

* דירוג: לפי מיקום דירוג הכשירויות כך שלא מתחשבים בגודל הפער בין כשירויות לעומת שיטת הרולטה בה יש משמעות לגודל הפער בין הערכים. היתרון הוא שמצטמצם הסיכוי לסחף גנטי והתכנסות מוקדמת כאשר יש פער ערכי כשירות גבוהים בין הפרטים באוכלוסייה. כך לא מתייחסים לערך הפער אלא לדירוג הקבוע.
* טורניר: כל פעם שני הורים נבחרים באקראית ומתחרים על מקום בקבוצה ממנה ייבחרו ההורים. בין שני ההורים הפרט המנצח הוא זה עם הכשירות הגבוהה יותר. החסרון הוא שזה יכול להאט את ההתכנסות בגלל שפתרונות טובים יכולים להתחרות זה מול זה בסיכוי שווה לעומת שפתרון טוב בהכרח יתחרה מול פתרון גרוע. כך פתרונות טובים מסויימים שהפסידו לא יעמידו צאצאים ופתרונות גרועים כן יוכלו להעמיד צאצאים אם ניצחו מול פתרון גרוע מהם. היתרון הוא שנשמר גיוון גבוה וצמצום האפשרות לסחף גנטי.
* מצב מתמיד: רוב הפריטים ממשיכים לדור הבא ורק הפריטים בעלי ערכי הכשירויות הנמוכים ביותר מתחלפים ע"י אחת משיטות הסלקציה האחרות שציינתי. חסרון- סחף גנטי לכיוון ערכי הכשירויות הגבוהים ביותר. יכול להיות שפתרון מוצלח נמצא דווקא בשילובים עם הפתרונות הגרועים, וניתן לפספס אותם כי מחליפים רק את הגרועים. יתרון – פתרונות מוצלחים נשמרים (אליטיזם).

ד. שחלוף יתבצע ע"י החלפת ביט בודד כלשהו מתוך ה10 בין ההורים בצורה אקראית.

ה. החישוב ייעצר לאחר 100 דורות.

ו. גודל האוכלוסייה ההתחלתית היא 50 והאוכלוסייה הראשונית תבחר בצורה אקראית.

השפעת גודל האוכלוסייה על ההתכנסות: ככל שתבחר אוכלוסייה גדולה יותר כך תהיה התכנסות מאוחרת יותר, סיכוי לבדיקת אפשרויות רחב יותר וכך גם סיכוי גדול יותר למציאת פתרון מינימום טוב יותר. החסרון הוא זמני חישוב וריצה ארוכים. אוכלוסייה קטנה תעלה את הסיכוי להתכנסות מוקדמת עקב מחסור במגוון אופציות, סיכוי גבוה יותר לסחף גנטי וכיסוי נמוך של מרחב החיפוש ובכך מציאת פתרון פחות טוב. עם זאת, היתרון הוא זמן ריצה/חישוב קצר.

1. התכנסות מוקדמת תתכן כאשר פרט בעל ציון כשירות גבוה באוכלוסייה "ישתלט" מוקדם מידי על שאר הפתרונות האפשריים באוכלוסייה בגלל סיכויי רבייה גדולים יותר. כלומר, מדובר בצמצום מוקדם מידי של מגוון האפשרויות ובכך פגיעה באפשרות למציאת פתרון אופטימלי יותר שיכל להימצא אם היו נבדקים יותר אופציות מגוונות.

גישת הנישה מקטינה את הסיכוי להתכנסות מוקדמת מכיוון שבמקום מציאת גנום אופטימלי ע"י אוכלוסייה אחת, בגישה זו מחלקים את האוכלוסייה לתתי-אוכלוסייה וכך גם אם נמצא בתת-אוכלוסייה כלשהי סכימה דומיננטית "שתשתלט" על שאר האופציות, עדיין יתבצע שיחלוף בין תתי-אוכלוסיות שונות כל כמה דורות. שחלוף בין תתי-אוכלוסיות יכניס אופציות חדשות לכל תת-אוכלוסייה. ככה בו בזמן אפשר להתכנס אל פתרון אופטימלי מקומי ללא פיזור גדול מידי, וגם לשמור על מגוון האפשרויות לשחלוף ע"י ערבוב תתי-האוכלוסיות.

**האלגוריתם**

השתמשתי במטריצה ריבועית בעלת x שורות על x עמודות על מנת לייצג את המסלול, כאשר נקודת ההתחלה היא 0,0 ונקודת הסיום היא x-1,x-1.

כמות המכשולים מיוצגת על ידי אחוז שטח המכשולים מתוך כלל שטח הריבוע. כל מכשול מיוצג על ידי תא שנבחר אקראית באיתחול הסביבה.

ייצוג כרומוזום נעשה ע"י מערך של מספרים שלמים מתוך אופציות {0,1,2,3}. כל מספר מייצג כיוון תנועה, כך שהמסלול מתחיל מנקודת הפתיחה ומתקדם על גבי תאי המטריצה לפי הכיוון הנבחר בכל אלל. גודל הכרומוזום מחושב דינמית לפי מספר התאים במטריצה כיוון שזהו מקסימום התנועות שניתן לבצע.

תחילה יוגדר לכל ניסוי גודל המטריצה, אחוז המכשולים וגודל האוכלוסייה. מתוך ערכים אלה יחושבו מספר המכשולים במטריצה, וכמות ההורים שיעברו לדור הבא בתהליך האליטיזם.

לאחר שהסביבה מאותחלת מחושבת אוכלוסיה אקראית התחלתית וערכי הכשירות של הפרטים בה.

הדורות הבאים מחושבים בלולאה בתוכה נעשה חישוב הדור הבא לפי הדור הקודם (בהתחשב במסלול הנוכחי וערכי הכשירות) וגם מופעלות פונקציית מוטציה על הדור החדש ופוקנציית חישוב ערכי הכשירות החדשים. בסוף האיטרציה נעשת בדיקת התכנסות.

פונקציית חישוב הדור הבא: שלב ראשון- אליטיזם שיישמר פתרונות טובים מהדור הקודם. לפי הקבוע שהחלטתי שומר את הX איברים עם הכשירות הכי גבוהה ומכניסים אותם לדור הבא כמו שהם. לאחר מכן מכניסים גם X פריטים רנדומליים בנוסף כדי שיישמר גיוון. כלומר כרגע ללא זיווג העברנו X2 לדור הבא כמו שהם.

שלב שני- זיווג. השתמשתי בשיטה משולבת של דירוג ורולטה. הרי הערכים בינהם משווים הם המנה של ערך הכשירות של פרט בסכום ערכי הכשירות באוכלוסייה. קודם עבור צאצא מספר i בוחרים i קבוצת הורים אפשריים שהם במיקומים 0 עד i במערך במערך הממויין לפי הערכים. מתוך קבוצה זו נגריל הורה ראשון בשיטת הרולטה. את ההורה השני נגריל בצורת הרולטה מתוך כל המערך. כלומר הורה ראשון נבחר מתוך קבוצה מצומצמת של i הכי טובים לפי מיקום דירוג, בשיטת רולטה לתת-קבוצה זו, והורה שני נבחר אקראית מתוך כל האוכלוסייה לפי שיטת הרולטה.

בזיווג בודקים מי ההורה עם המסלול הארוך יותר ולוקחים את רוב המסלול של הורה זה. ככה נשמר חלק גדול תקין של מסלול והוא יותר דומה להורה עם המסלול הארוך. את השינוי המהותי בצאצא עושים בקצוות – בהם מגרילים כל גן לפי שני ההורים בצורה אקראית בערך. כך בעצם בוחנים כיוונים חדשים ללא וויתור על תקינות המסלול.

פונקציית הכשירות: מקבלת את הכרומוזום והסביבה ומחשבת את הכשירות על ידי חלוקת 1 בציון פונקציית ניקוד העונשים. בפונקציית הניקוד "מענישים" כרומוזום במספר מקרים (ככל שניקוד יותר נמוך-פתרון יותר טוב). בודקים החל מנקודת ההתחלה את מספר התנועות התקינות לפני שנתקעים בקיר, במכשול או במסלול מעגלי (מכיוון שאנחנו מחפשים מסלול כמה שיותר קצר). אם מגיעים לפתרון עד הסוף, אז "מענישים" ככל שהמסלול ארוך. אם לא מגיעים לפתרון, מענישים בהתאם למרחק נקודת העצירה מנקודת הסיום ובהתאם למספר הצעדים שנעשו עד העצירה. כאן שיחקתי עם ערכים קבועים שונים לחומרת העונש בניסויים שונים עד שראיתי שיש התכנסות מאוזנת. לאחר מספר ניסויים הסקתי שהאלגוריתם עובד יותר טוב ככל שמעודדים את הפתרונות לנוע במסלולים שרחוקים מהקצוות וגם שההתקדמות היא בשני הצירים במקביל – ולכן הוספתי עונש על מצבים בהם פער ערך x שונה משמעותית מערך y.

פונקציית המוטציה: מקבלת מערך ממויין של אוכלוסייה לפי ערכי כשירות, ערכים של סיכוי למוטציה בכרומוזום וסיכוי למוטציה בגן. מגרילים צאצאים לפי הסיכוי למוטציה בכרומוזום לפי ערך ששיחקתי איתו לגבי כמות האוכלוסיה הכי מוצלחת שתשאר כמו שהיא (אליטיזם) מחליטים אם להעביר את הכרומוזום מוטציה או לא. אם הכרומוזום נבחר למוטציה – עבור כל גן מגרילים (לפי הסיכוי למוטציה בגן) האם יעבור מוטציה בנקודה זו ואם כן מגרילים כיוון חדש.

פונקצית בדיקת ההתכנסות: בודקים אם יש כרומוזום המייצג מסלול תקין עד לסוף המטריצה, ואם כן בודקים האם היה שיפור בערך הכשירות של הפתרון בפתרון בדור הבא תוך כדי שסופרים ע"י COUNTER את מספר האיטרציות שעברו מאז שנמצא פתרון ולא היה שיפור בSCORE. כך אנחנו נותנים הזדמנות נוספת למציאת אפשרות פתרון טובה למרות שהייתה התכנסות.

תנאי עצירה, שתי אפשרויות:

- ע"י בדיקת פונקציית הכשירות

- אם המסלול לא התכנס – 50 דורות מקסימום.

**תוצאות**

הרצת האלגוריתם נעשתה עבור:

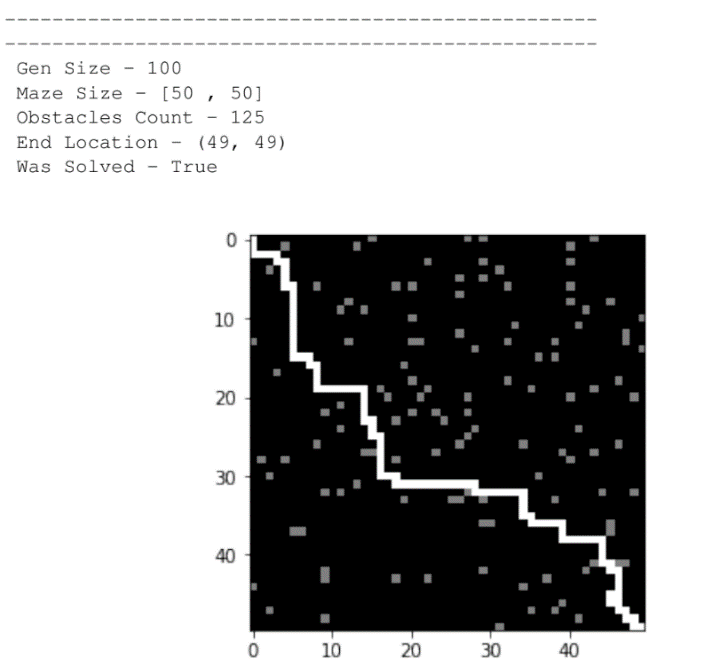
3 גדלי אוכלוסייה שונים: 25, 50, 100

3 גדלי מטריצה: 10X10, 25X25, 50X50

3 כמויות מכשולים: 0% , 5%, 10%

**בעמודים 9-38 יש פירוט חזותי של תוצאות כל ניסוי**

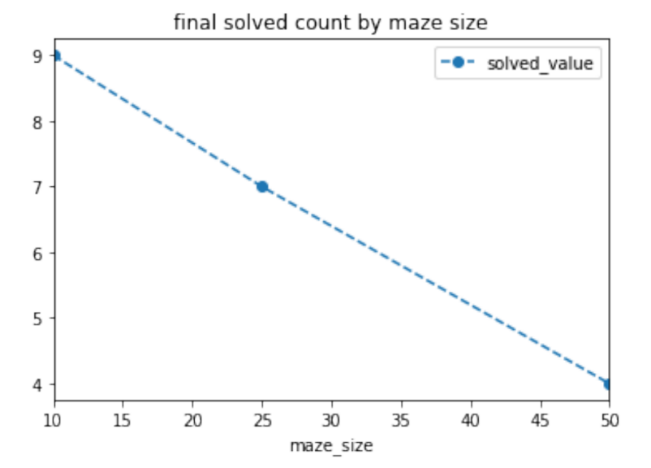
לדוגמא: ריצת ניסוי שהצליחה להתכנס לפתרון עם גודל אוכלוסייה 100, גודל מטריצה 50 ומספר מכשולים 125. תאי המסלול מסומנים בלבן ותאי המכשולים באפור.



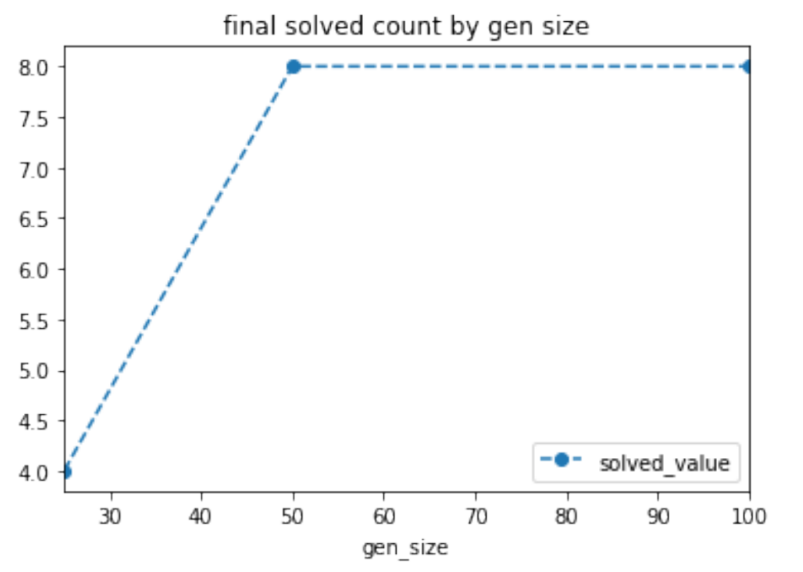
**בעמוד 39-44 נמצאים הגרפים בהם כל פעם בדקתי איך משתנה מסויים משפיע על ריצת האלגוריתם.**

המסקנות הכלליות הן:

* ככל שכמות המכשולים גדולה יותר כך פוחת הסיכוי להתכנסות לפתרון וזאת משום שהאלגוריתם לא יכול לחזות מראש היכן יש מכשול וככל שיש יותר מכשולים אז יש יותר הפרעות במסלול שצריך להתגבר עליהם על ידי שחלוף ומוטציות.
* אם נסתכל על הגרף המציג את מספר הניסויים שהתכנסו לפתרון כתלות בגודל המטריצה, נראה כי ישנה תלות ליניארית בין המשתנים וככל שגודל המטריצה קטן יותר כך סיכוי גבוה יותר להתכנס לפיתרון. זאת משום שיש פחות תאים לבדיקה ופחות קומבינציות שצריך לחשב.



* במבט על הגרף המציג את מספר הניסויים שהתכנסו לפתרון כתלות בגודל האוכלוסייה נראה כי לאחר עלייה ליניארית ישנה הגעה לרוויה (פלאטו) בו הגדלת מספר האוכלוסייה לא משפיעה על ההתכנסות. אני מניחה שזה בגלל שהשתמשתי באלגוריתם בשיטת אליטיזם והעדפת הכשירים יותר לרבייה – כך שגם באוכלוסייה גדולה סכימות דומיננטיות ישתלטו על סכימות אחרות ויהיה צמצום במגוון.



ניתן לראות בעמודים הרלוונטיים גרפים נוספים שעוסקים בניתוח התוצאות.