‏קורס:

מודלים מתמטיים

שם המגיש: פח'ראלדין רזי

שם המרצה: פרופ' כתריאל חגי

בנדיט מרובה ידיים



|  |
| --- |
| **1.הקדמה** |
|  |
|  |
|  |
| בעבודה הזו אני אציג את הבעיה של בנדיט מרובה ידיים (Multi Armed Bandits), זו בעיה מתורת ההסתברות ששייכת לקבוצת בעיות שנקראות ""Sequential decision problem שבהם מקבלים החלטות באופן סדרתי בתנאי אי-וודאות, בעיה זו מתארת מצב שבו מהמר נכנס לקזינו ורוצה לשחק במכונת מזל בעלת מספר ידיים (N) (בד"כ למכונות המזל יש יד אחת ,בבעיה הזו נניח שיש לאותה מכונה מספר ידיים) ,ואחרי שמושכים אחת מהידיים המכונה פולטת סכום כסף,לפעמים פחות, לפעמים יותר, לפעמים בכלל לא.על המהמר לבחור איזו יד למשוך ובאיזה סדר כך שימקסם את הרווח שלו על גבי מספר משחקים (משיכות) נתון (T).  הפעם הראשונה שהיה צורך להבין את הבעיה הזו הייתה ב1933 כאשר ויליאם תומפסון ניסה להשוות בין טיפולים שונים כאשר האפקטיביות של הטיפולים לא ידועה והציג לראשונה את האלגוריתם *Thompson sampling* שעד היום נחשב לאחד האלגוריתמים המוצלחים ביותר. [1]  אחרי שהתעלמו מהבעיה לזמן מה היום ניתן לראות שישנם הרבה מאמרים בנושא זה, ונחקר ע"י מתמטיקאים, מדעני מחשב ועוד. ראה [2] להמשך קריאה.  הבעיה הזו קיבלה הרבה תשומת לב מכיוון שהיא מתארת מצב שבו יש מתח בין החיפוש כדי למצוא את האפשרות הכי טובה (היד הכי רווחית),לבין ניצול האפשרות הכי רווחית עד כה(היד שפלטה הכי הרבה כסף עד כה), מצב זה מופיע בהרבה בעיות אמיתיות,וניתן למדל אותן באותה צורה, למשל:[5]   * *ניסוי קליני*: לזהות מתוך N תרופות שונות את התרופה היעילה ביותר. משחק = ניסוי. * האיזון בין הניצול לחיפוש הוא מאוד חשוב בבעיה הזו, כי מצד אחד נרצה לחקור תרופות חדשות שיכולות להיות יותר יעילות (חיפוש) ומצד שני נרצה לטפל בחולים בצורה אפקטיבית גם בזמן הניסוי (ניצול) . המתח בין ניצול וחיפוש ברור, אם נחליט לבחור בכל פעם את התרופה שאנחנו יודעים שהיא הכי יעילה עד כה נאבד את האפשרות לחקור תרופות חדשות שיכולות להיות הרבה יותר אפקטיביות וכך נפסיד בטווח הארוך. לעומת זאת אם נבחר בכל פעם באקראי מבין התרופות נוכל לחקור כמות יותר גדולה של תרופות אבל זה פוגע בחולים בטווח הקצר. ולכן חשוב מאוד לאזן בין השניים כך שמצד אחד תהיה לנו ההזדמנות לחקור תרופות אחרות תוך שמירה על הבריאות של החולים (מזעור הסיכונים). להסבר נוסף ראה [3].      * *שיווק באינטרנט*: לזהות מתוך N פרסומות שונות איזו מהן מושכת את העין הכי הרבה, "רווח" מתקבל כאשר מקליקים על המודעה. גם פה נרצה למצוא את הפרסומת הכי טובה אבל גם נרצה למקסם את החשיפה של המוצר ולכן גם פה נצטרך לנצל את הפרסומת הטובה עד כה. להסבר נוסף ראה [7]. * *מכירות באינטרנט*: (לשם פשטות נניח שמדובר במוצר שקיבלנו בחינם או יצרנו בעצמנו בלי עלויות) נרצה לקבוע איזה מחיר לגבות מהלקוחות כדי למקסם את הרווח הכולל ,לשם כך נבחר N מחירים שונים שנראים לנו הוגנים, מצד אחד נרצה לחפש את המחיר הכי רווחי ומצד שני נרצה לנצל את המחיר שעד כה היה כי רווחי. * ועוד הרבה *בעיות אחרות*: לבחור בין מספר קורסים,מספר מאמרים, או אפילו בין מספר אפשרויות לעריכת עבודה (כמו זאת ) , שבה "רווח" נוצר ע"י מידת ההבנה של הקורא.   בעבודה הזו אעשה מידול מתמטי לבעיה הנ"ל ואציג את האלגוריתם *epsilon greedy ,* אלגוריתםהיוריסטי ופשוט שפותר את הבעיה ואתמקד באיזון בין חיפוש לניצול*,* כדי לעשות זאת עשיתי סימולציות שמתארות את ההתנהגות של האלגוריתם ועניתי על השאלות הבאות:   1. האם האלגוריתם הזה יותר טוב מהאלגוריתם הטריוויאלי (אלגוריתם חמדן)? 2. איך לבחור את הפרמטר של האלגוריתם כדי לקבל ביצועים יותר טובים? ובמה זה תלוי? |
| **2. תיאור הבעיה והאלגוריתם אפסילון חמדן** |
| לפני שניגש לפתרון הבעיה אציג מספר הנחות ונקודות חשובות שיעזרו בניתוח ומידול הבעיה בהמשך. |
| ***2.1 הנחות המודל*** |
| להלן מספר הנחות שילוו אותנו בעבודה הזו:   1. יש N ידיים למכונת המזל . 2. נשחק את המשחק T פעמים, בכל פעם נמשוך את היד שנבחרה. 3. אחרי כל "משיכה" נקבל רווח מסוים (או הפסד) המוגרל מפונקצית התפלגות מסוימת. 4. כל "משיכה" מחזירה רווח אקראי, ולכל "יד" יש תוחלת קבועה אבל לא ידועה. 5. המטרה שלנו היא למקסם את הרווח הכולל שלנו אחרי ה- T משחקים. 6. אנחנו נלמד על כל יד ע"י משיכתה וצפייה ברווח שהתקבל. \*   (\*) אילו ידענו כמה כסף נרוויח בממוצע מכל יד הפתרון היה טריוויאלי , היינו בוחרים ביד בעלת התוחלת המקסימלית של כמות הכסף בכל פעם . ההנחה היא שאין לנו ידע מוקדם כזה. |
| ***2.2 המתח בין חיפוש וניצול*** |
| המתח הזה נוצר כי מצד אחד אנו רוצים לנצל את הידיים שהשיגו ביצועים טובים בעבר,ומצד שני נרצה לחפש ידיים אחרות שיכולות להיות יותר טובות.  כשמדברים על בעיית MAB אי אפשר להתעלם מהבעייתיות הזו, כל האלגוריתמים שקיימים מנסים לאזן בין חיפוש וניצול בגישות שונות, לפני שנתקדם ננסה להבין במה מדובר ע"י דוגמה:  נתאר לעצמנו מצב בו אנו נכנסים לקזינו ובוחרים אחת ממכונות המזל (הידיים) באקראי, ואחרי כל "משיכה" אנו מרוויחים 10$ בממוצע, זה מצב שנשמע כל כך טוב שאנו לא נחשוב על לנסות (לחפש) ידיים אחרות,  אבל אחרי כמה דקות נכנס שחקן לקזינו, מתיישב לידך, מושך את היד ומרוויח 1000000$ .  עכשיו בטח תחשוב , למה לא ניסיתי? .  לכן נרצה מצד אחד לנצל את המכונות (הידיים) שידוע לנו שהן רווחיות אבל לא לשכוח לחקור מכונות שיכולות להיות יותר רווחיות.  השאלה איך בוחרים את המכונות ובאיזה סדר כדי למקסם את הרווח הכולל? ישנם הרבה אלגורתימים שתוקפים את הבעיה בצורות שונות , אני אתמקד באלגוריתם שנקרא "אפסילון חמדן" ( ( *epsilon greedy* שהוא הפשוט מביניהם,למרות הפשטות של האלגוריתם הזה מסתבר שהוא מגיע לתוצאות יותר טובות מאלגוריתמים יותר מסובכים. ראה [6,3]. |
| ***2.3 האלגוריתמים*** |
| במהלך העבודה אחקור את האלגוריתם "אפסילון חמדן" ואשווה אותו לאלגוריתם הטריוויאלי "החמדן".   * + 1. **האלגוריתם החמדן**   זהו אלגוריתם מאוד פשוט שבו מקבל ההחלטה (המהמר) יבחר את היד שהייתה הכי רווחית\* עד כה ולא ינסה ידיים שאין לו מידע עליהם/ יש ידיים שידוע לו שהן רווחיות יותר, אם ישנם מספר ידיים עם רווחיות זהה (ומקסימלית) האלגוריתם יבחר ביניהם באקראי. למשל בהתחלה כשלמקבל ההחלטה אין שום ידע מוקדם – הוא יבחר באקראי אחת מהידיים.  (\*) יד רווחית = היד שבממוצע פלטה סכום כסף חיובי. הכי רווחית זו יד שבממוצע פלטה הכי הרבה כסף. לעומת זאת יד מפסידה היא יד שבממוצע פלטה סכום כסף שלילי (הימרנו והפסדנו).   * + 1. **האלגוריתם "אפסילון חמדן"**   זהו אלגוריתם פשוט אך קצת יותר מתחכם מהאלגוריתם החמדן, עכשיו אנחנו ברוב המצבים מתנהגים כמו האלגוריתם החמדן, אך לעיתים אנו בוחרים אחת מהידיים באקראי בשם החיפוש.  **נתאר את האלגוריתם מתמטית:**  *המדיניות של הבחירה:*  בהסתברות  נבחר במכונה שהייתה הכי רווחית עד כה ובהסתברות  נבחר באחת מ N הידיים באקראי.  כאשר  פרמטר המייצג את אחוז הפעמים שבהם נבחר לעשות חקירה(חיפוש).  \*נשים לב כי כאשר  נקבל את האלגוריתם החמדן.  נגדיר:  - תוחלת הכסף ש- ידa פולטת. "הערך האמיתי" שאותו המהמר לא יודע.  - ההערכה של המהמר אחרי t משחקים לגבי תוחלת הכסף שיד a פולטת .  - מספר הפעמים שמשכנו את היד a עד המשחק ה t.  אחרי ההגדרות הנ"ל נוכל לתאר בצורה מתמטית את:   1. *המדיניות:*   (1)  כאשר:  Rand – מספר אקראי בין 0 ו 1 שנלקח מהתפלגות יוניפורמית.(בכל בחירה מגרילים אותו מחדש)  Randomly – הכוונה שבחורים אחת הידיים באקראי עם התפלגות אחידה.   1. *הערכת התוחלת:*   כדי להעריך את התוחלת של יד a נחשב ממוצע על היסטוריית הרווחים, כלומר: (2)  כאשר . הרווחים שהתקבלו במשחקים בהם בחרנו את היד a.  *דוגמא מסכמת:*  נניח שיש לנו מכונה עם 2 ידיים , מהמר "חכם" שמכיר את האלגוריתם הזה התיישב מול המכונה ,איך הוא ישחק?  (בהנחה ש- )  *שלב 1:*  בהתחלה אין לו שום ידע לגבי רווחיות הידיים, לכן מבחינתו : , אז הוא יבחר אחת מהידיים באקראי . נניח שבחר ביד מספר 1, משך אותה והרוויח 3 שקלים. עכשיו אחרי שקיבל את המידע על יד 1 הוא מעדכן את ההערכה שלו:  .  *שלב 2*:  עכשיו הוא מסתכל על ההערכות שלו, ואז מתנהג לפי המדיניות של האלגוריתם, כלומר בהסתברות 0.8 יבחר את היד בעלת ההערכה הגבוהה ביותר (יד מספר 1) ובהסתברות 0.2 יבחר אחת מהידיים באקראי.  וכך הלאה..  כלומר, המהמר : יבחר את היד לפי המדיניות (1), יעדכן את ההערכות לפי (2). באופן אטרטיבי עד שיפסיק לשחק. |
| 1. **חקירת ההתנהגות של האלגוריתם אפסילון חמדן** |
| ***3.1 סימולציות*** |
| אחרי שהבנו איך האלגוריתם "אפיסלון חמדן" עובד ננסה להבין איך הוא מתנהג ובכמה הוא יותר טוב מהאלגוריתם ה "חמדן" . כדי לעשות זאת עשיתי מספר סימולציות וניתוחים אנליטיים:  בכל הסימולציות שיעשו בהמשך:   * תוחלת הרווח שכל יד פולטת נבחר מהתפלגות נורמלית סטנדרטית, כלומר ונתייחס לזה כמו קודם : "הערך האמיתי" של יד k. * הרווח שמתקבל אחרי כל משיכה של יד a מתפלג נורמלית עם תוחלת ששוה ל "ערך האמיתי" של אותה יד ושונות 1, כלומר : . * מספר הידיים של המכונה N=10. * אם מבצעים מספר ריצות אז אם לא נאמר אחרת, "הערך האמיתי" של כל יד מוגרל באופן אקראי בכל ריצה. |
| **3.2.1 סימולציה ראשונה** |
| ניתן לראות ש: , כי כאשר t שואף לאינסוף נקבל כי בחרנו כל יד אינסוף פעמים , כלומר  , לכל a , ולכן הממוצע שואף לתוחלת. (החוק החזק של מספרים גדולים)  את הסימולציה הראשונה עשיתי כדי להראות את העובדה הנ"ל.  כדי לעשות זאת בחרתי מספר משחקים גדול שידמה את המצב ( t שואף לאינסוף ).  בחרתי את הפרמטרים של הסימולציה כך :  התוצאות מוצגות בגרף הבא :  **גרף 1.1**  C:\Users\Razi\Desktop\h.jpg  הסברים ומסקנות:   * בגרף הזה רואים את השינוי של  כפונקציה של . עבור אחת הידיים. * ניתן לראות כי . אפילו עבור היד המפסידה (ראה 2.3.1). * את היד הזו מקבל ההחלטה בוחר רק 100 פעמים מתוך 10,000 המשחקים מכיוון שהיא פולטת סכומי כסף שליליים. * אפשר לראות שזה נכון גם לגבי שאר הידיים בגרף 1.2.   **גרף 1.2**  C:\Users\Razi\Desktop\9f.jpg  הסברים ומסקנות:   * כמו הגרף הקודם, ציר X מתאר את מספר הפעמים שהיד נבחרה. * ניתן לראות כי היד הרווחית ביותר (הגרף בפינה הימנית העליונה ) נבחרה ביותר מ 80% מהמשחקים * ניתן לראות כי ככל ש  גדול יותר הפער בין הערך  לערך מצטמצם כצפוי. |
| **3.2.2 סימולציה שנייה** |
| את הסימולציה הבאה אני מציג כדי להראות את תהליך הלמידה של האלגוריתם.  עשיתי 4 ריצות זהות כאשר "הערך האמיתי" של הידיים נשאר קבוע (כלומר עבור יד a נבחר  פעם אחת ולא השתנה מריצה לריצה), ובכל סימולציה: .  *להלן התוצאות:*  **גרף 2.1**  4plots  *הסברים ומסקנות:*   * הערכים בגרף חושבו כך:   אחרי כל משחק חושב ממוצע הרווחים בכל המשחקים עד כה, למשל עבור t=5 חושב הממוצע כך,  **,** כאשר reward(t) זה הרווח שהתקבל במשחק t.   * ניתן לראות כי יש הבדלים ברורים בין הסימולציות בעלי אותם פרמטרים. זה נובע מהאקראיות של הבחירה (כאשר האלגוריתם "מחפש") וגם בגלל האקראיות של הרווח. * כפי שניתן לראות בהתחלה שהעקומות עולות בצורה ברורה ואז בחלק מהעקומות יש ירידה קטנה,   העלייה בהתחלה ברורה מכיוון שהמהמר מתחיל לחקור עד שהוא מוצא יד עם רווח חיובי.את הירידה שמופיעה בחלק מהגרפים אפשר להסביר בגלל העובדה שאנחנו עושים ממוצע של הרווחים עד לאותה נקודת זמן ולכן כל סכום שלילי ישפיע מאוד על הממוצע, סכום שלילי או "הפסד" יכול לנבוע משני מצבים:   1. בזמן חקירה, המהמר מושך יד עם תוחלת שלילית ומפסיד סכום יחסית גדול. 2. אקראיות של הרווח, יתכן מצב שאנו משחקים עם יד שתוחלתה חיובית ונפסיד מדי פעם, מספיק שזה יקרה מספר קטן של פעמים והממוצע יקטן בצורה ברורה.  * למרות ההבדלים אפשר להבחין בכך שהאלגוריתם "לומד" ומגיע לתוצאות טובות יותר עם הזמן, אבל מהירות הלמידה יכולה להשתנות מסימולציה וסימולציה ולכן ניתן לראות שמקבלים רווח ממוצע שונה. |
| **3.2.3 סימולציה שלישית** |
| אנו מבינים שהשונות ברווח הממוצע נובעת משני גורמים עיקריים:   1. שונות בערכים האמיתיים של המכונות. 2. האקראיות בתהליך הלמידה.   כדי להבין את ההשפעה של כל גורם על השונות נעשו שתי סימולציות :   1. מכונות משתנות בין הריצות. 2. אותן מכונות עבור כל הריצות.   כאשר בכל סימולציה:  1000 משחקים,3000 ריצות ועבור כל ריצה חושב הרווח הממוצע.כאשר .  *להלן התוצאות:*   1. *ערך אמיתי קבוע:*   **גרף 3.1**  C:\Users\Razi\Desktop\untitled.jpg   1. *ערך אמיתי משתנה:*   **גרף 3.2**  C:\Users\Razi\Desktop\untitled.jpg  *מסקנות:*   * ניתן לראות שהשונות הנובעת מתהליך הלמידה של האלגוריתם לא גבוהה במיוחד )סטיית תקן מדגמית=0.1) * לעומת זאת השונות כאשר גם הערך האמיתי של המכונות מוגרל באקראי בכל ריצה גדולה הרבה יותר (סטיית תקן מדגמית=0.5) * מה שכן ניתן לומר זה שברוב המקרים נקבל רווח ממוצע גבוה יחסית לתוחלת הרווח של הידיים   שנבחרה מהתפלגות נורמלית עם תוחלת 0 כמו שצוין קודם. |
| * + 1. **סימולציה רביעית** |
| אחרי שראינו שהשונות היא גדולה בין הריצות:  עשיתי את הסימולציה הבאה כדי להראות איך האלגוריתם מתנהג "בממוצע".   * המיצוע נעשה על גבי 2000 ריצות.   *להלן התוצאות:*  **גרף 4**  הרווח הממוצע בכל משיכה כתלות במספר המשחקים    *מסקנות:*   * הגרף הזה מציג את הרווח הממוצע (על 2000 ריצות) אחרי כל משחק. * מתאר מצב שבו השחקן "חמדן" ובכל פעם בוחר את היד הכי רווחית עד כה,בלי לחקור ידיים אחרות. * ניתן לראות כי אנחנו נרוויח יותר כאשר ושכאשר  נמצא את היד הכי רווחית מהר יותר אבל נבחר בה רק ב 91% מהמקרים . (). * ניתן לראות כי הגרף מתייצב על ערך מסוים של רווח ממוצע.(עבור אפסילון נתון). |
| * + 1. **סימולציה חמישית** |
| בסעיף זה:   1. אחשב אנליטית את תוחלת הרווח המתקבל אחרי הרבה זמן. 2. אשווה את התוצאה האנליטית לתוצאות סימולציה.   שלב 1: פתרון אנליטי  בכדי לחשב את תוחלת הרווח, נסתכל על זה כך:   * אחרי מספר גדול של משחקים האלגוריתם אמור לדעת איזה יד הכי רווחית.   (לפי העובדה ש  לכל a).  ולכן מקבל ההחלטה יבחר ב  מהמקרים את היד הכי רווחית , וב  אחת הידיים באקראי.  בשפה מתמטית:  ,  כאשר:  -  התוחלת של המקסימום מבין 10 משתנים מקריים נורמלים.  -  תוחלת הממוצע של 10 משתנים מקריים נורמלים . (תוחלת 0, שונות 1)  - ברור כי  (תוחלת הסכום שווה לסכום התוחלות).  לכן : .  כלומר ככל שאפסילון יותר גדול נקבל כי הרווח הממוצע מכל משיכה קטן.(בטווח הארוך)  לכן מספיק לחשב את , או במילים אחרות את התוחלת של המקסימום,נעשה זאת כך:   1. מהי פונקציית ההתפלגות המצטברת של המקסימום?     \*משתנים מקריים בלתי תלויים.  כאשר F זו פונקצית ההתפלגות המצטברת של מ"מ נורמלי סטנדרטי.   1. חישוב התוחלת:     כדי לחשב את זה נעזרתי במטלב [8], והתשובה המתקבלת .  ולכן האלגוריתם ירוויח בממוצע :  , נבדוק זאת עבור שני ערכים:   * , רווח ממוצע צפוי : * , רווח ממוצע צפוי:   שלב 2: סימולציה  כדי לבדוק זאת נבנה סימולציה שבה T=3000 :  *להלן התוצאות:*  **גרף 5**  plotre   * ניתן להבחין בכך שהרווח הממוצע שואף לערכים שציפינו להם (1.38,1.52). * אומנם עבור  עדיין קצת רחוק מ 1.52 וזה מכיוון שככל ש  קטן יותר לוקח לאלגוריתם יותר זמן להתכנס ליד הרווחית ביותר. |
| * + 1. **סימולציה שישית** |
| הסימולציה הבאה מראה את המתח בין חיפוש וניצול בצורה יותר ברורה:   * נעשו 1000 ריצות . * הרווח הממוצע חושב עבור כל הזוגות  בגרף, כאשר: ונבדק בקפיצות של 0.02. * כל ערך בגרף חושב ע"י מוצע על כל ה 1000 ריצות.   להלן התוצאות:  **גרף 6.1**C:\Users\Razi\MATHwork\Machine learning\n armed bandit\newplot.jpg  C:\Users\Razi\AppData\Local\Microsoft\Windows\INetCache\Content.Word\untitled.jpg **גרף 6.2**  הסבר ומסקנות:   * גרפים אלה מראים את ממוצע הרווח שמקבלים מכל משיכה כתלות באפסילון. * ניתן לראות כי ככל שנגדיל את מספר המשחקים נעדיף אפסילון קטן יותר, זה מראה איך האלגוריתם מאזן בין "חיפוש" ו"ניצול" . * עבור כל עקומה, נחפש איזה אפסילון מחזיר לנו את הרווח הממוצע הגבוה ביותר,   כדי לעשות זאת נסתכל באיזה טווח מתקבל המקסימום :  עבור מספר משחקים = 2000 רואים שהמקסימום מתקבל כאשר  עבור מספר משחקים = 1000 רואים שהמקסימום מתקבל כאשר  עבור מספר משחקים = 500 רואים שהמקסימום מתקבל כאשר  עבור מספר משחקים = 300 רואים שהמקסימום מתקבל כאשר  עבור מספר משחקים = 100 רואים שהמקסימום מתקבל כאשר |
| **4. סיכום ומסקנות** |
| אחרי הניתוח שנעשה ברור לנו שהאלגוריתם "אפסילון-חמדן" מגיע לתוצאות הרבה יותר טובות מהאלגוריתם "החמדן", ככל שנשחק יותר משחקים ההפרש בין הרווח הממוצע של שני האלגוריתמים הולך וגדל עד לערך מסוים מכיוון שראינו כי בטווח הארוך האלגוריתם "אפסילון-חמדן" מוצא את היד הכי רווחית ובוחר אותה בכל פעם (בהסתברות ) .  ראינו כי בחירת האפסילון תלויה במספר המשחקים שאנו מתכננים לשחק , ככל שמספר המשחקים גדל נבחר אפסילון קטן יותר, למרות שבעבודה ניתנו המלצות לגבי הטווח שבחירת האפסילון צריך לזכור שמדובר בממוצע ולא בהמלצה "בטוחה" , הכל יכול לקרות בהסתברות מסוימת.  ההמלצות הנ"ל ניתנות על בסיס הנחות המודל שהוצגו, ולכן אינן תקפות עבור בעיה דומה עם הנחות שונות, למשל:   1. הנחתי שהרווחים מתקבלים מהתפלגות נורמלית, ניתן לבדוק מה קורה כאשר מדובר בהתפלגות אחרת. 2. הנחתי שהשונות זהה עבור כל המכונות ושווה ל1 לאורך כל העבודה, מה קורה כאשר השונות גדולה יותר או קטנה יותר? או כאשר כל מכונה מקבלת שונות אחרת? 3. כדי לבדוק איזה אפסילון לבחור עבור T מסוים עשיתי מיצוע על גבי הרבה ריצות כשבכל ריצה משחקים עם מכונות אחרות ("ערך אמיתי" משתנה) , ניתן לבדוק מה קורה כאשר הסימולציות זהות (המהמר נכנס לאותו קזינו כל יום ומשחק T משחקים)   תחת שינוי כלשהו יתקבלו המלצות אחרות לגבי בחירת האפסילון.  ראינו חסרון ברור באלגוריתם בטווח הארוך שלמרות שהוא יודע איזה יד הכי רווחית הוא עדיין "חוקר" ידיים אחרות וזה משהו שהיינו רוצים להימנע ממנו. הייתי ממליץ לשנות את האלגוריתם כך שאחרי מספר משחקים מסוים (כתלות באפסילון שנבחר) נפסיק את החיפוש ונתמקד בניצול. לחילופין אפשר לחקור את האלגוריתם (epsilon-decreasing) שהוא דומה מאוד לאלגוריתם הזה רק שהוא מקטין את האפסילון עם הזמן, כך שבהתחלה האלגוריתם נוטה "לחקור" ואחרי הרבה משחקים הוא נוטה "לנצל" , ניתן לחקור את האלגוריתם הזה ולנסות להבין איך לבחור את הפרמטרים השונים, מאיזה אפסילון להתחיל? באיזה קצב נקטין את האפסילון?. האם זה תלוי במספר המשחקים?  או אולי את האלגוריתם epsilon-first שבו עושים חיפוש טהור (בוחרים באקראי בכל משחק) ב  המשחקים הראשונים ואז עושים ניצול טהור של היד שמקבל ההחלטה רואה כרווחית ביותר\* ב המשחקים שנותרו.  ניתן לחקור איך בוחרים את הפרמטר אפסילון? איך זה תלוי במספר המשחקים?  \*ראה 2.3.1 |
| 1. **רשימה ביבליוגרפית** |
| |  |  | | --- | --- | | W. R. Thompson. On the likelihood that one unknown probability exceeds another in view of the evidence of two samples. *Biometrika*, 25(3-4):285–294, 1933. | [1] | | [https://en.wikipedia.org/wiki/Thompson\_sampling#History](https://en.wikipedia.org/wiki/Thompson_sampling%23History) | [2] | | V. Kuleshov and D. Precup, Algorithms for the multi-armed bandit problem, Journal of Machine Learning Research, vol. 1, pp. 1-48, 2000. | [3] | | Richard S. Sutton, Andrew G. Barto-Reinforcement Learning\_ An Introduction-The MIT Press (1998) . | [4] | | <http://www.ece.ucdavis.edu/~qzhao/Zhao10SPAWC_slides.pdf> | [5] | | <http://stevehanov.ca/blog/index.php?id=132> | [6] | | <https://www.inverse.com/article/13762-how-the-multi-armed-bandit-determines-what-ads-and-stories-you-see-online> | [7] | | <https://github.com/RaziEldeen/>  MATLAB implementation for epsilon-greedy algorithm. | [8] | |  | |