לימוד במערכות מורכבות (049004)

גיליון תרגילים 4

הגשה: מומלץ 5/6, נדרש 13/6

תרגיל זה יש לבצע ולהגיש באופן אישי (לא בזוגות)

1. קריאה: עיינו במאמר הסקירה

Kaelbling, Littman and Moore, "Reinforcement Learning – a Survey", *Journal of Artificial Intelligence Research* 4, 1996, pp. 237-285.

סכמו בכחצי עמוד את הבעיות הפתוחות העיקריות לזמן כתיבת המאמר, כפי שעולות ממנו.

2. שימוש בתכונת ההתכנסות של מרטינגלים:

נתבונן במשפחת הסתברויות מעבר $P^\theta = \{p^\theta(s'|s)\}$ התלויות בפרמטר $\theta \in \Theta$. נסמן ב- θ_0 את ערכו האמיתי של הפרמטר (שאינו ידוע מראש). תהי $(s_k)_{k\geq 0}$ השרשרת המרקובית המתקבלת עבור פרמטר θ_0. נגדיר את פונקצית הסבירות:

$$L_n(\theta) \triangleq P^\theta\{s_0,\ldots s_n\} = p_0(s_0)\prod_{k=0}^{n-1} p^\theta(s_{k+1}|s_k)$$

ואת יחס הסבירות: $\Lambda_n(\theta) = L_n(\theta)/L_n(\theta_0)$.

הראו כי הסידרה $\{\Lambda_n(\theta)\}_{n\geq 0}$ הינה מרטינגל (לכל θ). הסיקו מכך את התכנסות $\Lambda_n(\theta)$ (כאשר $n\to\infty$).

הערה: פונקצית הסבירות משמשת להגדרת וחישוב משערך הסבירות המירבית (MLE, Maximum Likelihood Estimator) $\hat{\theta}_n = \arg\max_\theta L_n(\theta)$. התכנסות פונקצית הסבירות משמשת בניתוח ההתכנסות של משערך זה.

3. תרגיל סימולציה: לימוד במשחק ״אבן נייר ומספריים״ (Rock-Paper-Scissors).

נתבונן בבעית הלימוד של אסטרטגיה אופטימאלית במשחק חוזר של המשחק הנ״ל, כאשר ידוע כי היריב מרקובי וסטציונרי. כלומר: בחירת הפעולה של היריב (שיכולה להיות רנדומלית) תלויה אך ורק בצמד הפעולות שנבחר בצעד הקודם.

א. הגדר את הבעיה כבעיית החלטה מרקובית. ניתן להניח קריטריון מהוון, עם מקדם היוון קרוב ל-1, ופונקציית רווח המקבלת ערכים (-1, 0 1+). מהם הפרמטרים הלא-ידועים? בחר מעתה ערכים כלשהם (לא סימטריים, לא דטרמיניסטיים ולא טריוויאליים) לפרמטרים אלה.

ב. השתמש בתכנות דינאמי כדי למצוא את פונקצית הערך האופטימאלית.

ג. השתמש בלימוד-Q כדי ללמוד את המדיניות האופטימאלית.

ג.1. הנח ראשית כי מדיניות השחקן הלומד הינה אקראית (בחירת כל פעולה בהסתברות שווה). בדוק סכמות לימוד עם הגבר דועך ועם הגבר קבוע, השווה. מעתה ניתן להתמקד באחת מהאפשרויות שבדקת.

ג.2. הנח עתה כי מדיניות השחקן הלומד הינה חמדנית ביחס לפונקצית Q המשוערכת. יש לבחור תנאי התחלה 0 לפונקציה זו. האם קיימת בעיה בהתכנסות? בדוק פתרונות אפשריים ע״י הוספת ערור (exploration), וע״י שינוי תנאי ההתחלה של Q.

ד. בחר סכמת לימוד מתאימה מסוג actor-critic, והדגם את פעולתה על הבעיה הנדונה. בדוק את המקרה שבו שני חלקי האלגוריתם פועלים בקצב דומה, ואת שני המקרים שבהם האחד מהיר מהשני.

הנחיות כלליות לסימולציות: יש להציג גרפים מייצגים, עם הסברים מלווים. יש לבחור גדלים להצגה אשר מדגימים את איכות המדיניות הנלמדת לאורך זמן ואת הניקוד הנצבר בפועל (או קצב הצבירה) כתלות בזמן.