

לימוד במערכות מורכבות (049004)**גיליון תרגילים 3 (הגשה: 3.5.11)**

1. א. נתון תהליך החלטה דטרמיניסטי עם קריטריון ההחזר המהוון, כאשר הבקרה נבחרת לפי מדיניות סטציונרית דטרמיניסטית π . הצע אלגוריתם לומד פשוט (בדומה לאלגוריתם לימוד-Q לסביבה דטרמיניסטית שהוצע בכיתה) לחישוב פונקציית הערך V^π . הוכח התכנסותו.

ב. האם האלגוריתם שהצעתם יתכנס כאשר המדיניות π אינה דטרמיניסטית? הסבר.
2. א. אלגוריתם TD למחיר מהוון (פרק 4 עמ' 11 ברשימות): השלם את הפיתוחים הבאים.

א. הוכח את הנוסחה עבור עדכון ℓ -step lookahead.

ב. פתח את המימוש של אלגוריתם TD(λ) בעזרת eligibility trace.
3. תרגיל סימולציה: לימוד במודל סטטי (Multi-armed bandit).

נתבונן במודל החלטה נטול מצב, עם שתי אפשרויות החלטה ביניהן יש לבחור בכל מצב (ניתן לראות תהליך זה כ-MDP בעל מצב אחד בלבד ושתי פעולות). התגמול המתקבל עבור כל פעולה הינו אקראי, ופילוגו תלוי אך ורק בפעולה הנבחרת. פילוג זה אינו ידוע מלכתחילה.

לצורך התרגיל בחר את פילוגי תגמול כך שהפרש הממוצעים בין שתי הפעולות יהיה קטן יחסית לשונות. הנח פונקציית תגמול מהוונת עם מקדם היוון 0.95.

א. ישם את אלגוריתם TD(0) להערכת הערך V עבור מדיניות הבחירת בין הפעולות בהסתברות שווה. בדוק התכנסות עבור בחירות שונות של הגבר האלגוריתם (כולל הגבר קבוע ודועך), והשווה. בדוק גם את אלגוריתם TD(λ) עבור מספר ערכי λ .

ב. ישם את אלגוריתם SARSA להערכת פונקציית ערך-הפעולה Q עבור המדיניות הנ"ל, ובדוק התכנסות.

מעטה נבחר הגבר קבוע אך קטן לאלגוריתם השערוך.

ג. ישם אלגוריתם SARSA עם מדיניות בחירה חמדנית של הפעולות. האם אלגוריתם זה מתכנס תמיד לפעולה המיטבית?

ד. חזור על הסעיף הקודם עם מדיניות ϵ -חמדנית, ועם soft-max. בדוק עבור מספר ערכי פרמטרים והשווה.

ה. חזור על סעיף ג עם אחת מהאפשרויות לשינוי הדרגתי של המדיניות (החל מהמדיניות של סעיף א).