Sapienza
Università di Roma

Reward shaping in RL for $\mathrm{LTL}_f/\mathrm{LDL}_f$ Goals: Theory and Practice

Facoltà di Ingegneria dell'Informazione, Informatica e Statistica
Corso di Laurea Magistrale in Master in Engineering in Computer Science

Candidate
Marrco Favorito
ID number 1609890

Thesis Advisor
Prof. Giuseppe De Giacomo

Co-Advisor
Dr. co-advisor

Academic Year 2017/2018

Thesis defended on 1
in front of a Board of Examiners composed by:

Prof. ... (chairman)

Prof. ...

Prof. ...

Prof. ...

Prof. ...

Prof. ...

Prof. ...

Reward shaping in RL for $\mathrm{LTL}_f/\mathrm{LDL}_f$ **Goals: Theory and Practice**
Master thesis. Sapienza – University of Rome

This thesis has been typeset by LaTeX and the Sapthesis class.

Author's email: favorito.1609890@studenti.uniroma1.it

Abstract

write your abstract here

Contents

Chapter 1

Introduction

Chapter 2

Preliminaries

In this chapter we describe the background knowledge required for this work. We introduce Markov Decision Process (MDP) and Non-Markovian Reward Decision Process (NMRDP), common formalisms in the context of Reinforcement Learning. We describe Linear Temporal Logic over finite traces (LTL_f) and Linear Dynamic Logic over finite traces (LDL_f), that we use for define temporal goal in a RL setting. Then, we describe an important result about RL for NMRDP with $\text{LTL}_f/\text{LDL}_f$ rewards, that is the basis for this work.

2.1 Reinforcement Learning

Reinforcement Learning (Sutton and Barto, 1998) is a sort of optimization problem where an *agent* interacts with an *environment* and obtains a *reward* for each action he chooses and the new observed state. The task is to maximize a numerical reward signal obtained after each action during the interaction with the environment. The agent does not know a priori how the environment works (i.e. the effects of his actions), but he can make observations in order to know the new state and the reward. Hence, learning is made in a *trial-and-error* fashion. Moreover, it is worth to notice that in many situation reward might not been affected only from the last action but from an indefinite number of previous action. In other words, the reward can be *delayed*, i.e. the agent should be able to foresee the effect of his actions in terms of future expected reward.

In the next subsections we introduce some of the classical mathematical frameworks for RL: Markov Decision Process (MDP) and Non-Markovian Reward Decision Process (NMRDP).

2.1.1 MDP

A Markov Decision Process $\mathcal{M}$ is a tuple $\langle S, A, T, R, \gamma \rangle$ containing a set of *states* S, a set of *actions* A, a *transition function* $T : S \times A \to Prob(S)$ that returns for every pair state-action a probability distribution over the states, a *reward function* $R : S \times A \times S \to \mathbb{R}$ that returns the reward received by the agent when he performs action a in s and transitions in s', and a *discount factor* γ, with $0 \leq \gamma \leq 1$, that indicates the present value of future rewards.

2.1.2 Temporal Difference Learning

2.1.3 NMRDP

2.1.4 Reward Shaping

2.2 LTL$_f$ and LDL$_f$

2.2.1 Linear Temporal Logic for finite traces: LTL$_f$

2.2.2 Linear Dynamic Logic for finite traces: LDL$_f$

2.2.3 LTL$_f$ and LDL$_f$ translation to automata

2.3 RL for NMRDP with LTL$_f$/LDL$_f$ rewards

Chapter 3

RL for $\mathrm{LTL}_f/\mathrm{LDL}_f$ Goals

Chapter 4

Automata-based Reward shaping

Chapter 5

Experiments

Chapter 6

Conclusions

Appendix A

FLLOAT

Appendix B

RLTG

Bibliography

Richard S. Sutton and Andrew G. Barto. *Introduction to Reinforcement Learning*. MIT Press, Cambridge, MA, USA, 1st edition, 1998. ISBN 0262193981.