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   Co-Advisor

Prof. Giuseppe De Giacomo   Dr. co-advisor

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 $\text{LTL}_f/\text{LDL}_f$ 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 ($\textsc{ltl}_f$) and Linear Dynamic Logic over finite traces ($\textsc{ldl}_f$), that we use for define temporal goal in a RL setting. Then, we describe an important result about RL for NMRDP with $\textsc{ltl}_f/\textsc{ldl}_f$ 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.

A *policy* $\rho : S \to A$ for an MDP $\mathcal{M}$ is a mapping from states to actions, and represents a solution for $\mathcal{M}$. Given a sequence of rewards $R_{t+1}, R_{t+2}, \ldots, R_T$, the

expected return G_t at time step t is defined as:

$$G_t := \sum_{t=k+1}^{T} \gamma^{k-t-1} R_k \qquad (2.1)$$

where can be $T = \infty$ and $\gamma = 1$ (but not both).

The *value function* of a state s, the *state-value function $v_\rho(s)$* is defined as the expected return when starting in s and following policy ρ, i.e.:

$$v_\rho(s) := \mathbb{E}_\rho[G_t|S_t = s], \forall s \in S \qquad (2.2)$$

Similarly, we define q_ρ, the *action-value function for policy ρ*, as:

$$q_\rho(s, a) := \mathbb{E}_\rho[G_t|S_t = s, A_t = a], , \forall s \in S, \forall a \in A \qquad (2.3)$$

Notice that we can rewrite 2.2 and 2.3 recursively, yielding the *Bellman equations*:

$$v_\rho(s) = \sum_{s'} P(s'|s, a)[R(s, a, s') + \gamma v(s')] \qquad (2.4)$$

where we used the definition of the transition function:

$$T(s, a, s') = P(s'|s, a) \qquad (2.5)$$

We define the *optimal state-value function* and the *optimal action-value function* as follows:

$$v^*(s) := \max_\rho v_\rho(s), \forall s \in S \qquad (2.6)$$

$$q^*(s, a) := \max_\rho q_\rho(s, a), \forall s \in S, \forall a \in A \qquad (2.7)$$

Notice that with 2.6 and 2.7 we can show the correlation between $v_\rho^*(s)$ and $q_\rho^*(s, a)$:

$$q^*(s, a) = \mathbb{E}_\rho[R_{t+1} + \gamma v_\rho^*(S_{t+1})|S_t = s, A_t = a] \qquad (2.8)$$

We can define a partial order over policies using value functions, i.e. $\forall s \in S.\rho \geq \rho' \iff v_\rho(s) \geq v_{\rho'}(s)$. An *optimal policy ρ^** is a policy such that $\rho^* \geq \rho$ for all ρ.

2.1.2 Temporal Difference Learning

Temporal difference learning (TD) refers to a class of model-free reinforcement learning methods which learn by bootstrapping from the current estimate of the value function. These methods sample from the environment, like Monte Carlo methods, and perform updates based on current estimates, like dynamic programming methods.

Q-Learning and Sarsa are such a methods. They updates $Q(s, a)$, i.e. the estimation of $q^*(s, a)$ at each transition $(s, a) \rightarrow (s', r)$. Q-Learning uses the following update rule:

$$Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma \max_a' Q(s', a') - Q(s, a)] \qquad (2.9)$$

while Sarsa:

$$Q(s, a) \leftarrow Q(s, a) + \alpha[r + \gamma Q(s', a') - Q(s, a)] \qquad (2.10)$$

2.1.3   NMRDP

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

Chapter 3

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

Chapter 4

Automata-based Reward shaping

4.1   Reward Shaping

4.2   Reward shaping over automaton φ

Chapter 5

Experiments

5.1 BREAKOUT

5.2 SAPIENTINO

5.3 MINECRAFT

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.