Using Temporal Difference Learning to Blend Robot Behaviours

Author 1 and **Author 2**
Anonymous Submission 217
for ICAPS 2017

Abstract

Typically robots may be required to achieve multiple different objectives simultaneously. In this paper we concentrate on the common problem of a robot trying to navigate through a dense obstacle field. The robot is required to continuously move in the desired direction while avoiding collisions with the obstacles. We explore the idea of blending together two behaviours designed for the two different objectives. Eventually we learn a single unified policy for the complete task whose output is the amount of blending at each timestep. We use a receding horizon deliberative planner with a precomputed trajectory library to move towards the goal. Since the best trajectory selected by this planner might collide with obstacles, we blend this planner with a reactive behaviour designed for obstacle avoidance. We learn the amount of blending over several episodes and the resulting blended policy is able to outperform the naive deliberative planner for the given task.

Introduction

Robots often operate in partially known environments with densely cluttered obstacles (Daftry et al. 2016; Richter, Vega-Brown, and Roy 2015; Selekwa et al. 2008; Watterson and Kumar 2015; Bekris and Kavraki 2007). Usually in such environments the global goal is to reach a desired location or continuously move in a desired direction, and there is an obvious implicit sub-goal of avoiding collisions with the obstacles. We address the problem of simultaneously avoiding collisions with obstacles while moving towards the goal by learning a blended policy between deliberative goal-directed planning and reactive obstacle avoidance. In the past this problem has been solved for the case where full information about the environment is available a priori using sampling based methods (LaValle and Kuffner 2001), state lattice graphs (Pivtoraiko and Kelly 2005), and even hybrid robot behaviours (Wang, Yong, and Ang 2002). We cannot use these approaches since in our case the robot senses parts of the environment for the first time as it moves, thus a complete path from start to goal cannot be computed in advance. We must rely on a receding horizon approach to safely navigate the environment.

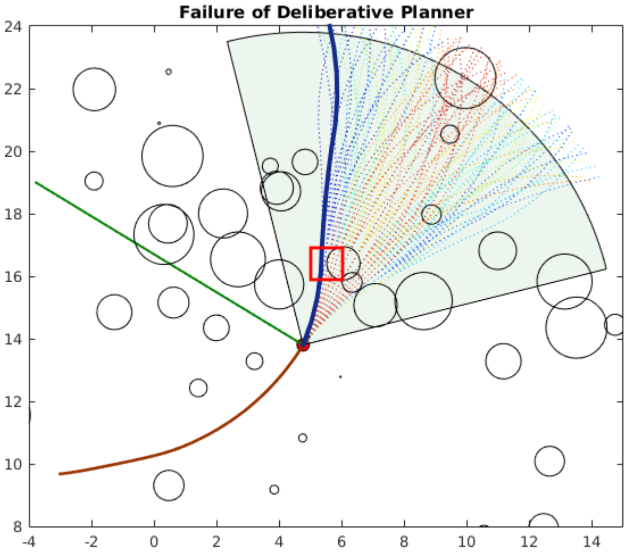

Figure 1: The best trajectory selected by the deliberative planner (solid blue curve) ends in collision with an obstacle (inside the red square), since it is influenced by heading in the goal direction (solid green line). Note: best seen in colour.

Our receding horizon deliberative planner utilizes a precomputed library of motion primitives. For every robot position in the environment and the surrounding obstacles, we cannot guarantee the existence of a collision free motion primitive in our library as shown in Figure 1. There are theoretical velocity and lattice resolution limits which can enable us to make such guarantees (Choudhury, Scherer, and Bagnell 2015), however we violate both these limits and thus an infinitely long collision-free path does not exist in the environment. This motivates the use of a reactive behaviour that is especially designed to avoid collisions with obstacles to augment the goal-directed behaviour of the deliberative planner. Our main contribution in this paper is learning a blended control policy that combines both the deliberative planner and the reactive behaviour via temporal difference reinforcement learning.

Previously, blending of robot behaviours has been stud-