Contents

Chapter 5

Exploration in MDPs

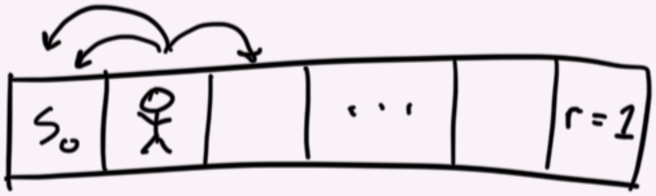
5.1 Introduction

In Chapter **??**, we explored algorithms for finding the optimal value and policy in an MDP when the transition and reward functions are unknown. However, we swept the issue of *exploration* under the hood. Namely, our algorithms might easily *overfit* to certain areas of the state space, missing out on possible better paths. This issue is especially relevant in **sparse reward** problems where reward might not be achieved until after many steps, and algorithms which do not *systematically* explore new states may entirely fail to learn anything meaningful.

For example, policy gradient algorithms require *signal* in the gradient to learn. In other words, if we never observe any reward, the gradient will always be zero, and the policy will never improve.

Example 5.1.1: Sparse Reward MDP

Here's a simple example of an MDP with sparse reward:



The agent starts in the leftmost state. There are three possible actions, two of which move the agent left and one which moves the agent right.

We also explored this issue in Chapter **??**