INTRODUCTION

Reinforcement learning gathers inputs and receives feedback by interacting with the external world. It outputs the best actions with that world. The goal is to maximize the total amount of rewards that it receives from taking actions in given states. In other words, the agent's goal is to follow a policy that maximizes the reward. Here, they investigate and compare some methods to select the best policy that gives maximum reward. They use prioritized sweeping algorithms.
Dyna style planning proceeds by generating imaginary experience from the world model and then applying model-free reinforcement learning algorithms to the imagined state transitions. They use the standard framework for reinforcement learning with linear function approximation.

Methods

They use four methods.
Algorithm 1 : Linear Dyna for policy evaluation, with random sampling and gradient-descent model learning
Algorithm 2 : Linear Dyna with PWMA prioritized sweeping (policy evaluation)
Algorithm 3 : Linear Dyna with MG prioritized sweeping (policy evaluation)
Algorithm 4: Linear Dyna with MG prioritized sweeping and TD(0) updates (control)

What is "prioritized sweeping"?
The states are waiting to be updated and they are kept in a queue. The queue is prioritized according to the size of their likely effect on the value function. high-priority states are popped off the queue and they are updated, it results to efficient sweeps of updates across the state space. There are two tabular prioritized sweeping algorithms in the literature. The first, due to Peng and Williams (1993) and to Moore and Atkeson (1993), is PWMA prioritized sweeping. The second form of prioritized sweeping, due to McMahan and Gordon (2005) is MG prioritized sweeping. There is a basic difference between these methods. PWMA puts the predecessors of every state encountered in real experience to the priority queue. It is not important that the encountered state was significantly changed. MG puts each encountered state on the queue, but not its predecessors. In this work, these algorithms update the value function from model-generated experience.

Limitations

The data here is from the model and the model is not a general system: it is deterministic and linear. These algorithms differ slightly from previous prioritized sweeping algorithms in that they update the value function from the real experiences and not just from model-generated experience. Thus they can perform a form of policy iteration—continually computing an approximation to the value function for the greedy policy.

Results

They make comparisons to model-free methods using variations of two standard test problems: Boyan Chain and Mountain Car. Prioritized sweeping gave in more efficient learning than simply updating features at random, and the MG method of prioritized sweeping seems to be better than the PWMA version. I will investigate later.