TExplore

Todd Hester and Peter Stone

TECHNISCHE UNIVERSITÄT DARMSTADT

Structure

TECHNISCHE UNIVERSITÄT DARMSTADT

1. Introduction
2. Related Work
3. Background
4. Algorithm
5. Experiments

Introduction

TECHNISCHE UNIVERSITÄT DARMSTADT

Goal

- Algorithm for real time targeted Exploration in large domains

General idea

- intelligent Exploration
- Sample efficient algorithm, which tries to explore a minimal number of states necessary to learn a good policy

Related Work

TECHNISCHE UNIVERSITÄT DARMSTADT

- There are a number of algorithms that address the exploration problem
- A few examples are R-MAX, SPITI, Bayesian RL methods or the Gaussian Process
- All have drawbacks, why they can‘t explore large domains in real time

Background – Markov Decision Process

TECHNISCHE UNIVERSITÄT DARMSTADT

- set of States S
- set of Actions A
- reward function R(s,a)
- transition function P(s‘|s,a)
- discrete state is represented by a vector of n discrete variables

$$s = < x_1, x_2, ..., x_n >$$

Background – Bellman equation

- $Q^*(s,a)$ of a state-action pair is an estimate of the future reward that can be obtained from (s,a)
- $Q^*(s,a)$ is determined by solving the Bellman equation :

$$Q^*(s,a) = R(s,a) + \gamma P(s' \mid s,a) \max_{a'} Q^*(s',a')$$

- Optimal policy $\pi(s) = \arg\max_a Q^*(s,a)$

Algorithm – Modul Learning

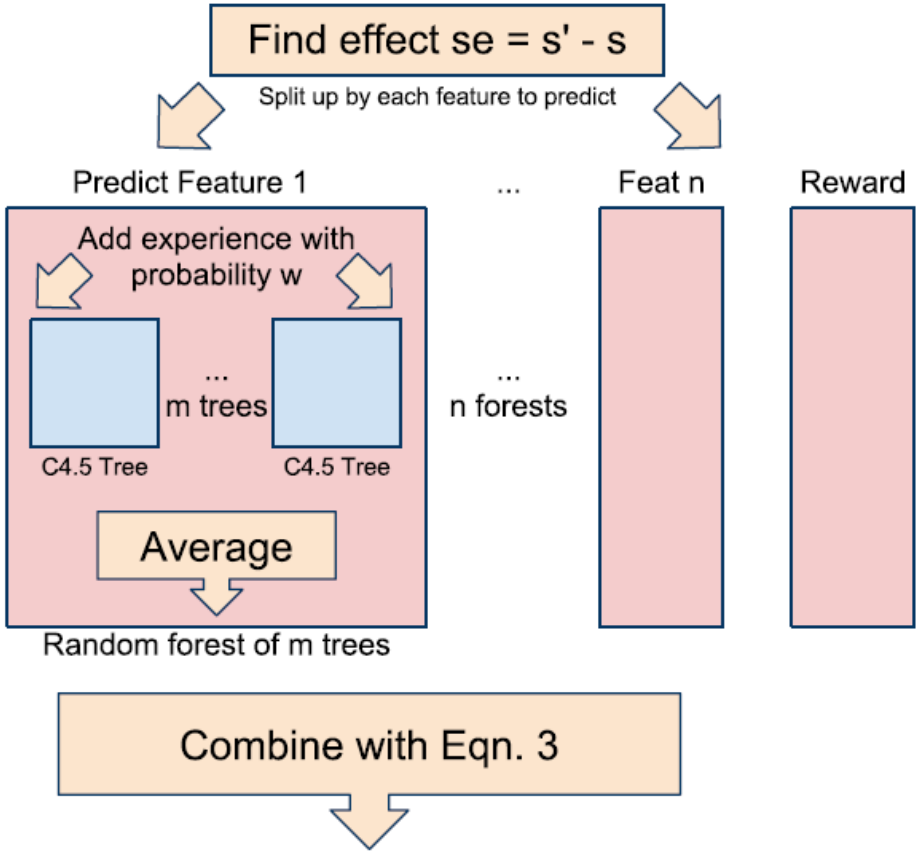
- Easier to generalize the transition effect (relative change) than the exact outcome across states → $s_e = s' - s$
- Learn seperate prediction of each of the n state features and put them together to a complete model
- The probability of the transition effect is the product of the probabilities of its n state features : $P(s_e \mid s, a) = \prod_{i=0}^{n} P(x_e^i \mid s, a)$

Algorithm – Modul Learning

TECHNISCHE UNIVERSITÄT DARMSTADT

- assumes that each state feature can be predicted independently
- Each model is build by using a random forest
  → each decision tree is trained on a subset of the agent experiences
- Final prediction is the average of all tree outcomes
- Each tree is build recursively using the C4.5 algorithm

Algorithm – Modul Learning



Algorithm - Exploration

TECHNISCHE UNIVERSITÄT DARMSTADT

- Agent is given instrinsic rewards based on the variance of the models predictions

$$R(s,a) = R_o(s,a) + b\sigma^2(s,a)$$

$$\sigma^2(s,a) = \frac{1}{n+1}\left[\sigma^2 R(s,a) + \sum_{i=1}^{n} \sigma^2 P(s_i^t \mid s,a)\right]$$

→ $R_o(s,a)$ is the original prediction of the model

→ $\sigma^2(s,a)$ is the variance in the model‘s predictions

Algorithm - Exploration

$$R(s,a) = R_0(s,a) + b\sigma^2(s,a)$$

- b is a coefficent either adding or subtracting intrinsic rewards from the model based on the variance
- b < 0, the agent will avoid uncertain states
- b > 0, the agent being driven to explore uncertain states
- b = 0, the agent will act greedily with respect to its model

Algorithm - Planning

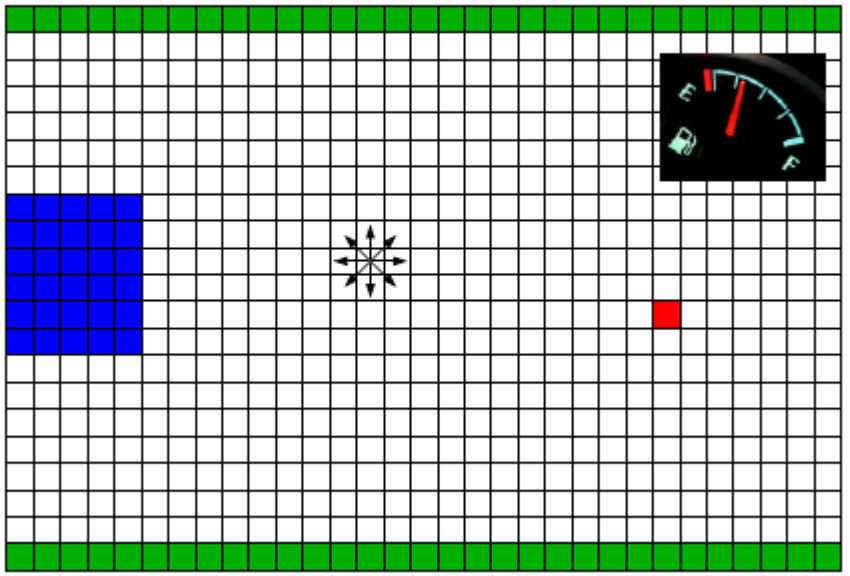
TECHNISCHE UNIVERSITÄT DARMSTADT

- Algorithm needs to replan what the best action is, by using the UTC algorithm
- Provide basic knowledge of the structure of the domain
- Seeding the agent with a few saple experiences from the domain

Experiments - Environment

- 21 x 31 cells
- 61 possible energy levels
- 8 possible actions

  → 317.688 state-actions
- Agent fuel level between 0 and 60
- State vector s is made up to three features: Row, Col and Fuel



- Fuel station reward $R(x) = base - (x \bmod 5)a$

  → x = column, base = reward for that row, a = variation of the costs across the columns

Experiments - Environment

TECHNISCHE UNIVERSITÄT DARMSTADT

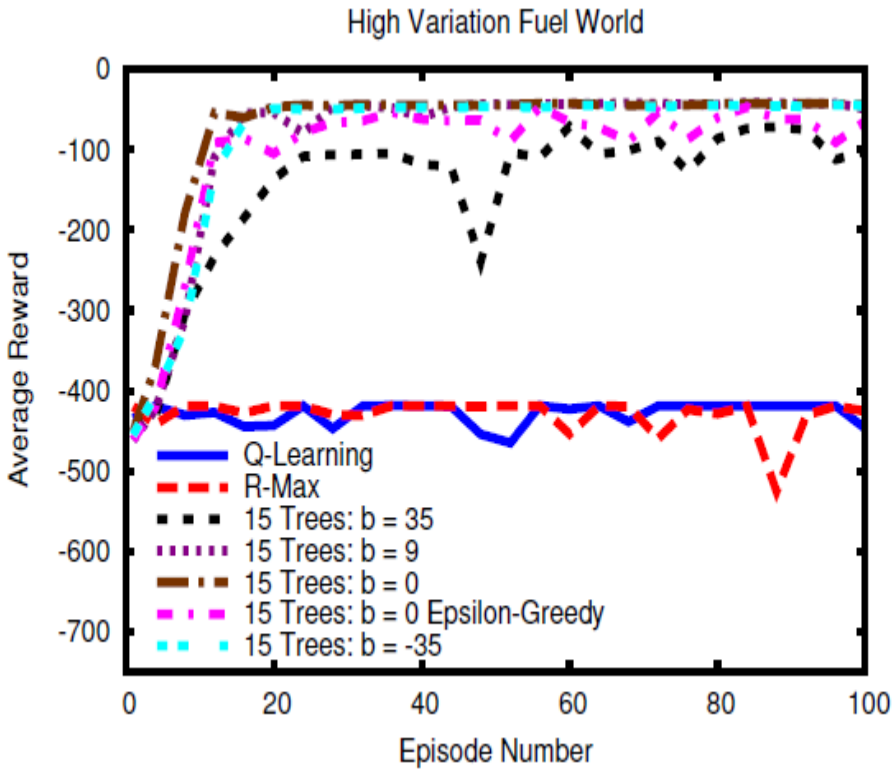
Low variation fuel world :

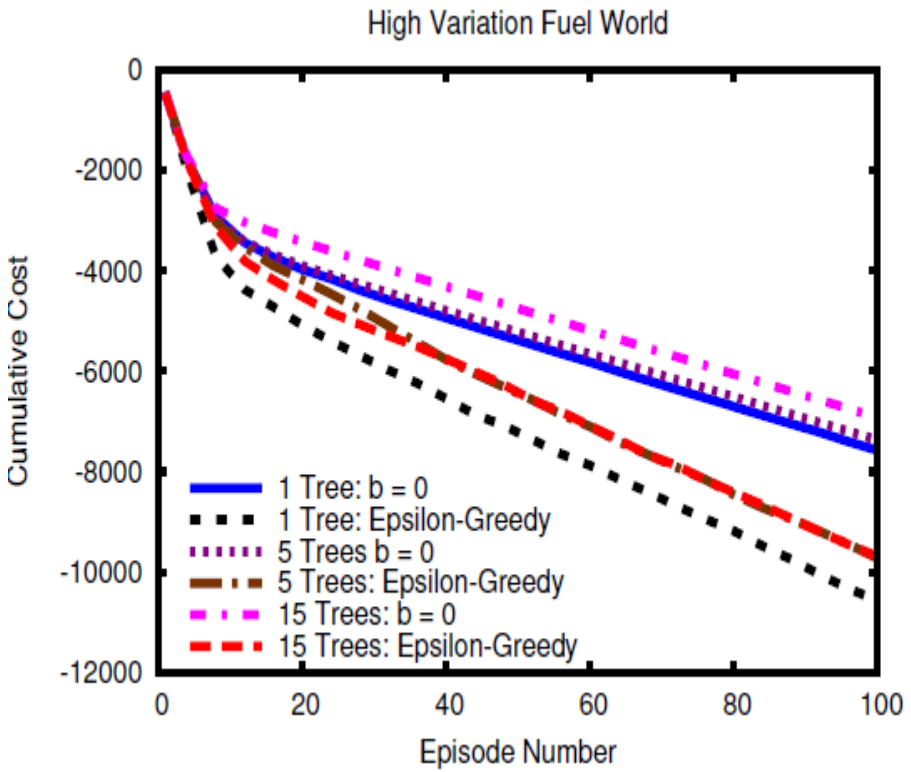
- a = 1.0, base = -18.0/-21.0 bottom/top row → reward = -18.0 to -25.0

High variation fuel world :

- a = 5.0, base = -10.0/-23.0 bottom/top row → reward = -10.0 to -33.0

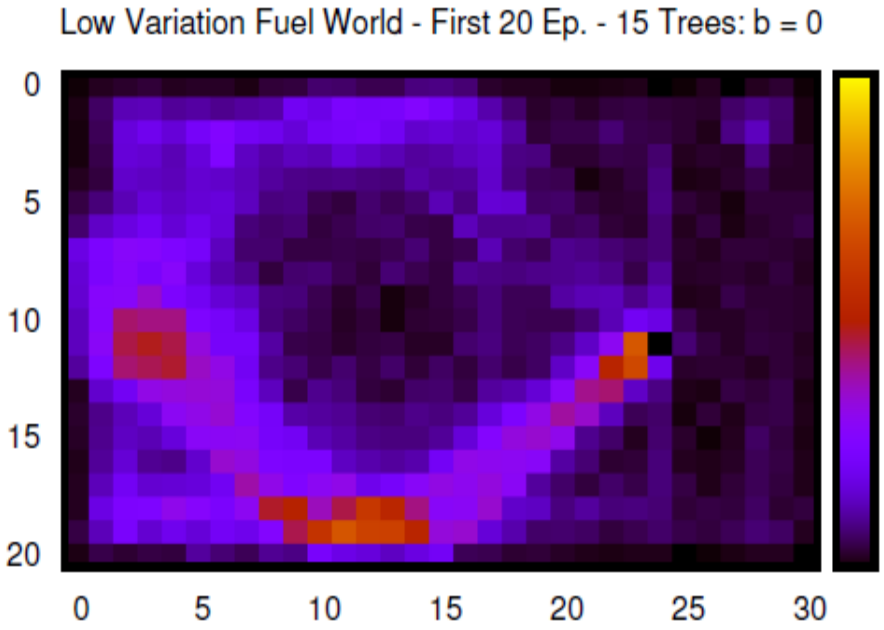
Experiments - Results

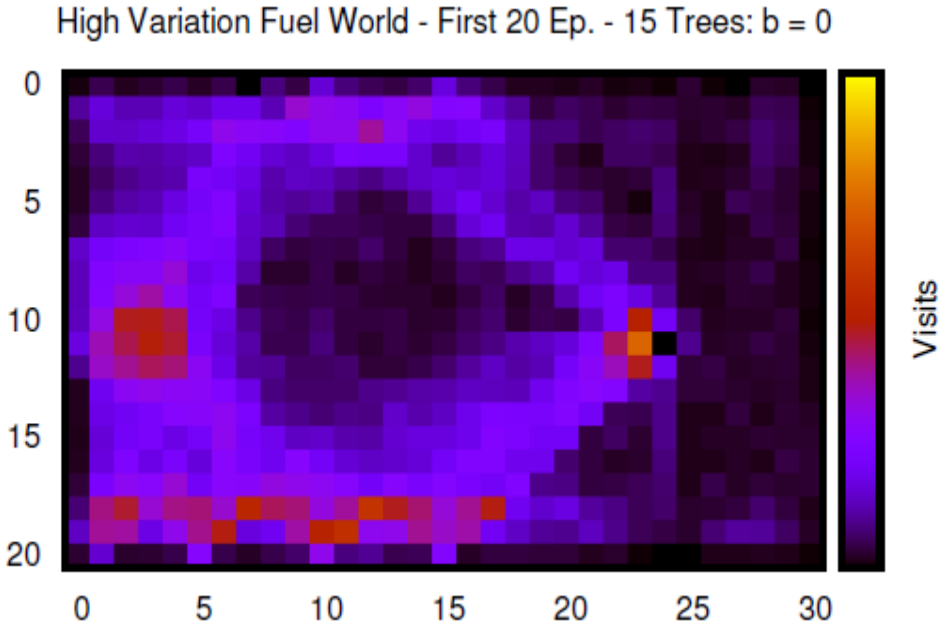




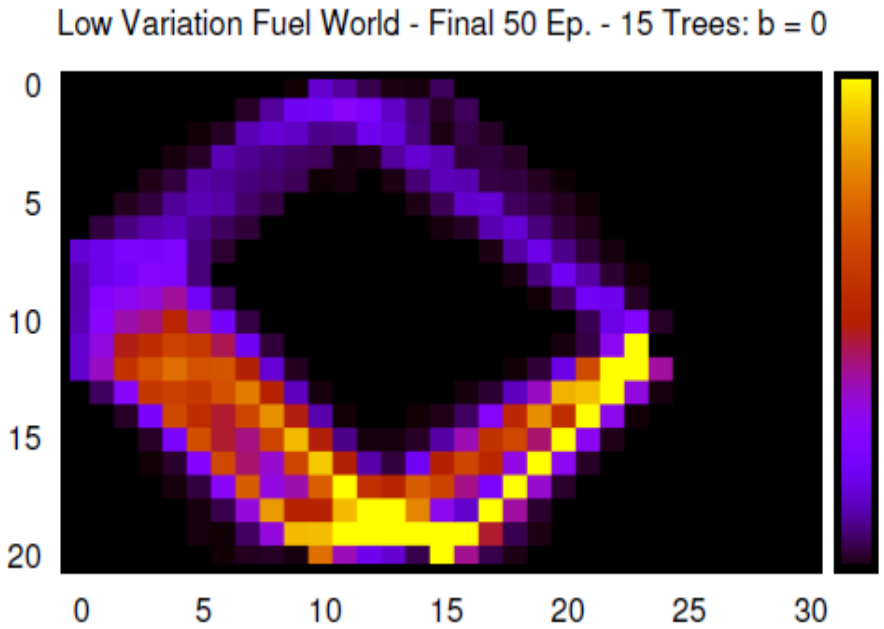
Experiments - Results

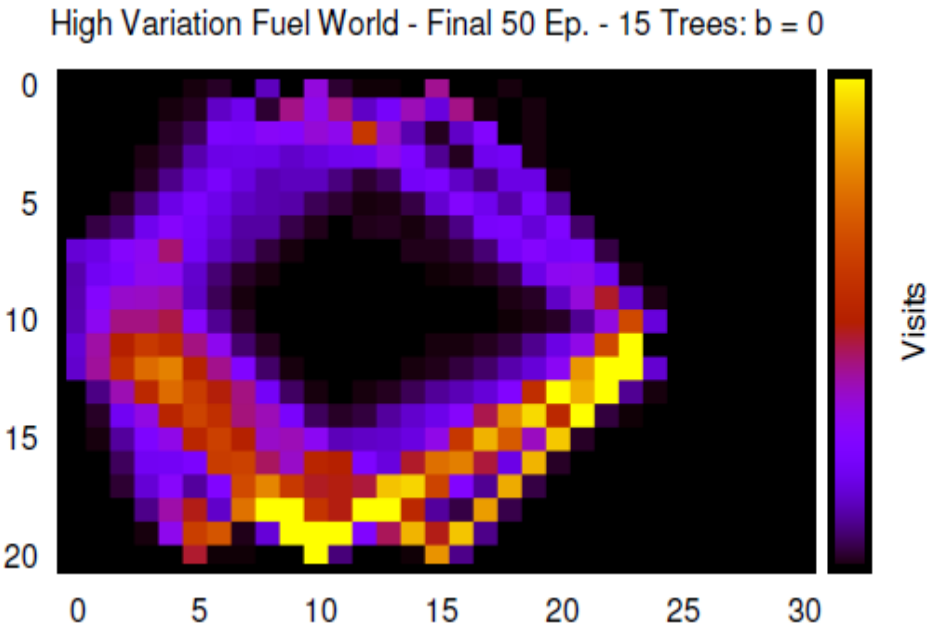
TECHNISCHE UNIVERSITÄT DARMSTADT





Experiments - Results





The End

TECHNISCHE UNIVERSITÄT DARMSTADT

Thank you for your Attention

Discussion

TECHNISCHE UNIVERSITÄT DARMSTADT

Any questions ?