Hierarchical Reinforcement Learning in the Taxicab Domain

Mitchell Keith Bloch
bazald@umich.edu

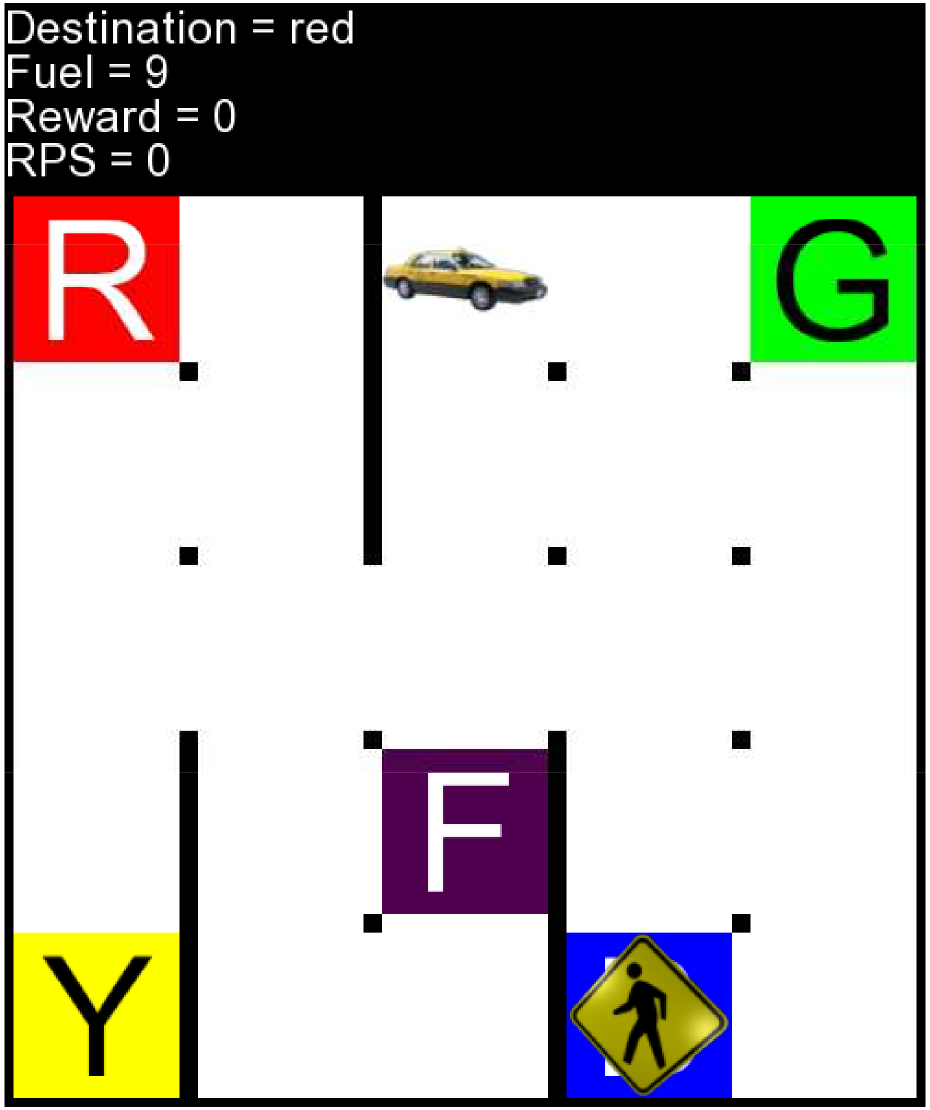
Soar Group
University of Michigan

Preview

- Taxicab Problem Domain
- RL and HRL
  - Dietterich's MaxQ Hierarchy
- Goals
- Agents
  - Performance
- Observations

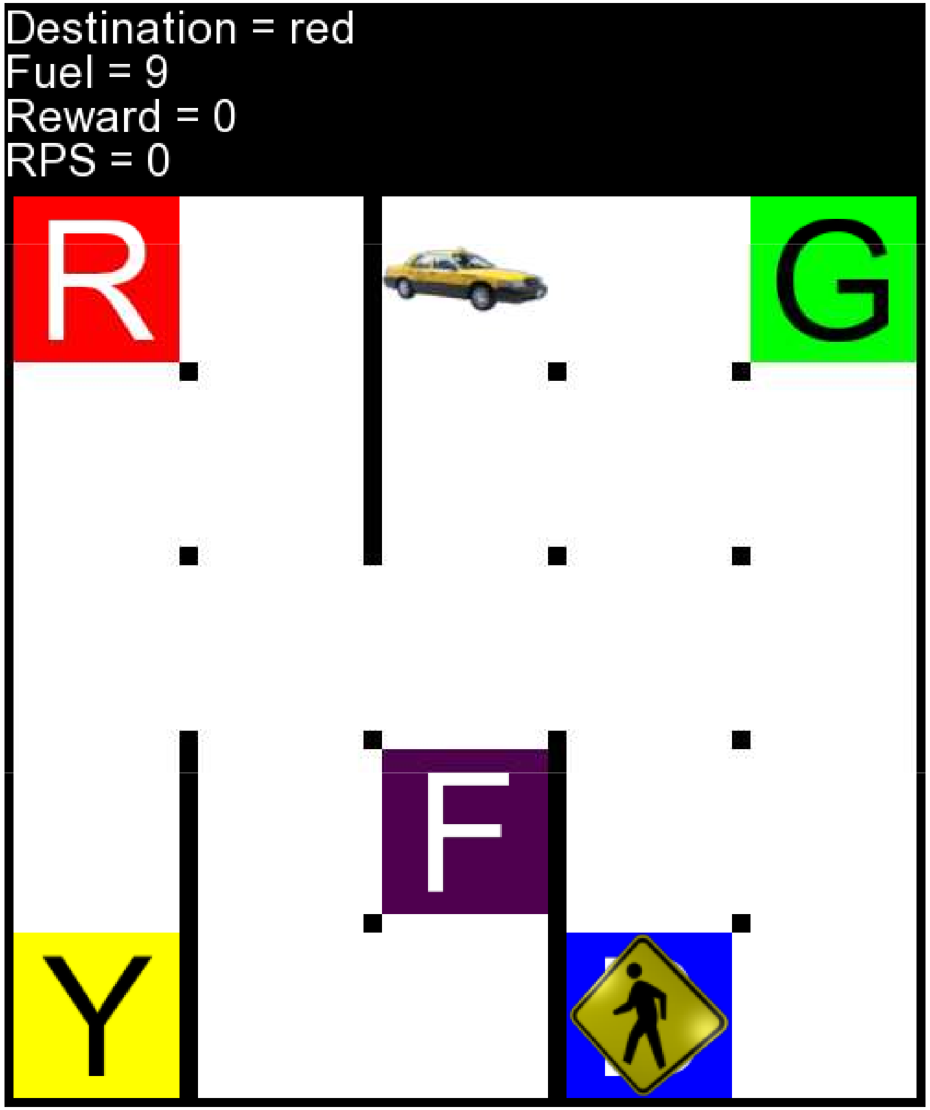
SML Taxicab Domain

- The agent's goal is to get the passenger and deliver it to the destination, without running out of fuel (in the Finite-Fuel Task)
- Given the fuel constraint, one false step can result in massive negative reward and incorrect learning



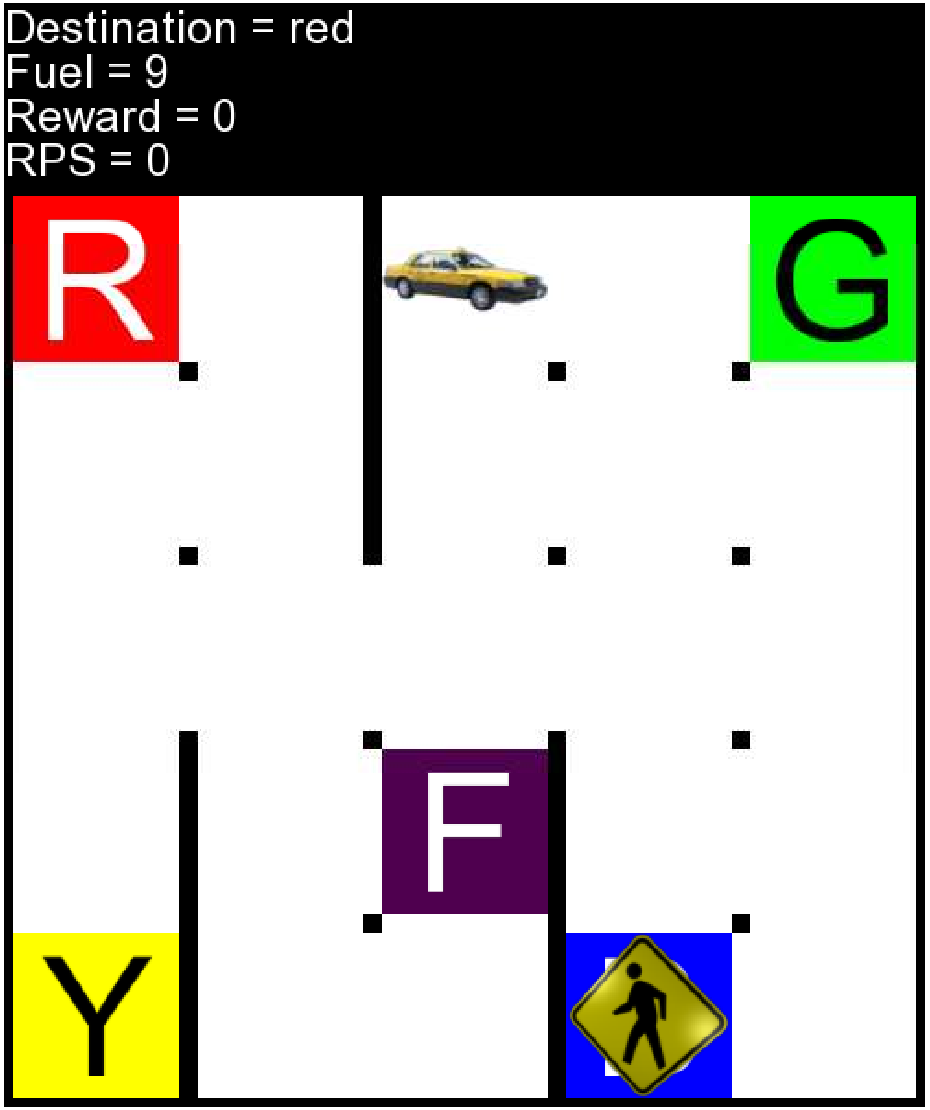
SML Taxicab Domain

- Taxi Starts Anywhere
- Passenger Starts at Red, Green, Blue, or Yellow
- Destination is Red, Green, Blue, or Yellow
- Fuel Initially Between 5 and 12 (inclusive)
- Maximum Fuel is 14



SML Taxicab Domain

- Actions are Discrete and Deterministic
  - Move North, South, East, or West
  - Pickup
  - Putdown
  - Refill
- In the Finite-Fuel Task, moving when out of fuel results in failure



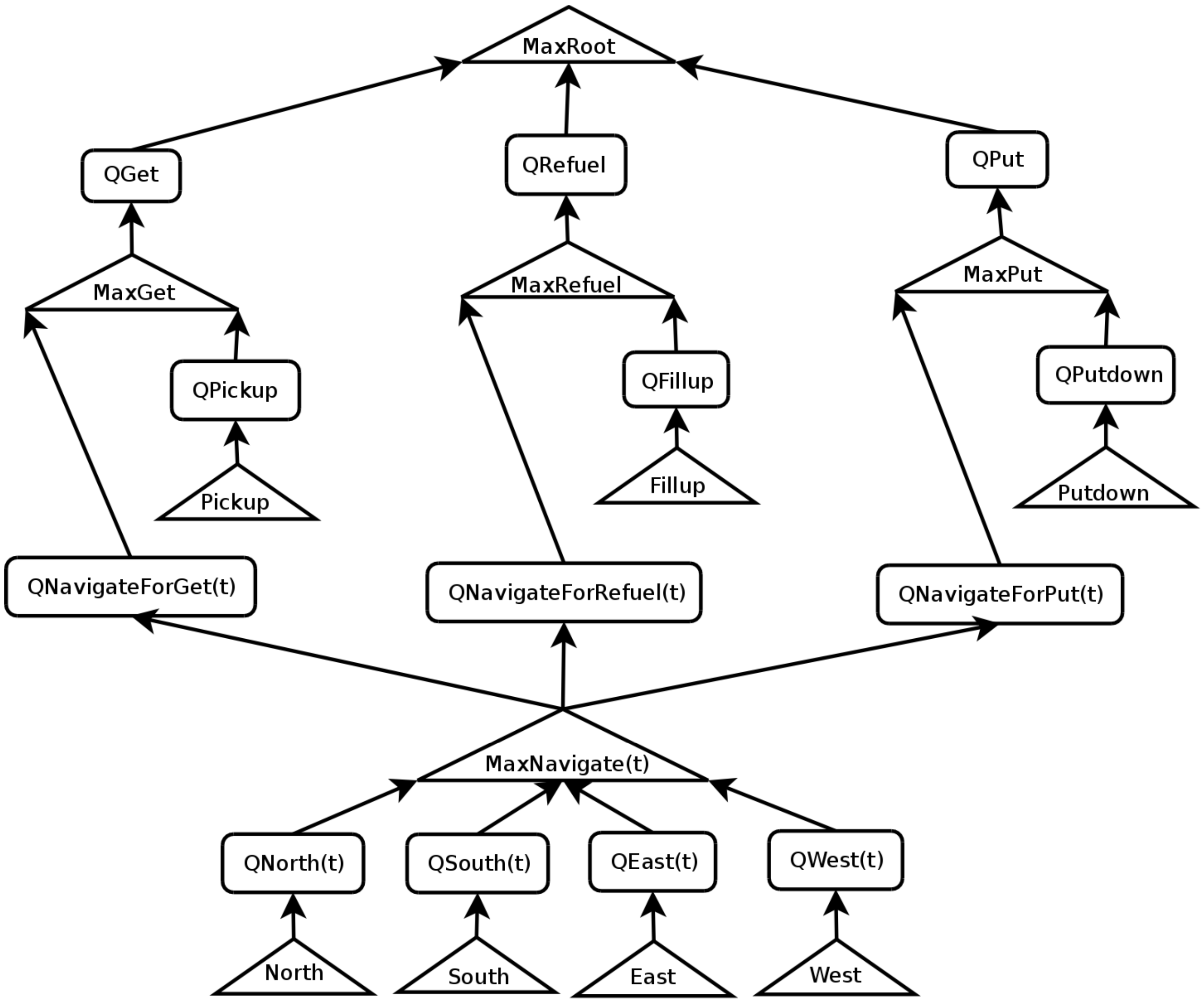
Reinforcement Learning

- RL agent must have a reward signal
  - One common evaluation metric is reward per step
- Agents typically learn a value function
  - Numeric indifferent preferences in Soar-RL
- Exploration policies vary significantly
  - Boltzmann indifferent selection biases exploration toward relatively promising actions
    - Important given the low probability of success when the fuel constraint is enforced

Hierarchical Reinforcement Learning

- Dietterich proposed MaxQ
  - Decompose task
    - Reduce the dimensionality of the problem
    - Enable transfer learning within the problem
  - Decompose reward signal
    - Subtasks receive reward for their decisions only
- Dietterich applied MaxQ to the Taxicab Problem Domain

Dietterich's Hierarchy



Goals

- Approximately reproduce Dietterich's work by applying MaxQ to the Taxicab Problem Domain in Soar-RL
  - Explore the capabilities of Soar-RL
  - Attempt to verify the original results

Soar-RL Parameters

- All my agents use
  - Learning Rate 0.3 (Dietterich used 1.0)
  - SARSA
  - Boltzmann Indifferent Selection
    - Initial Temperature 1.0 (Dietterich used 50.0)
    - Exponential Reduction Rate of 0.9999 (Dietterich varied this parameter at each MaxQ node)
    - Minimum Temperature of 0.05
  - No discounting
  - No eligibility traces

Four Task Variations

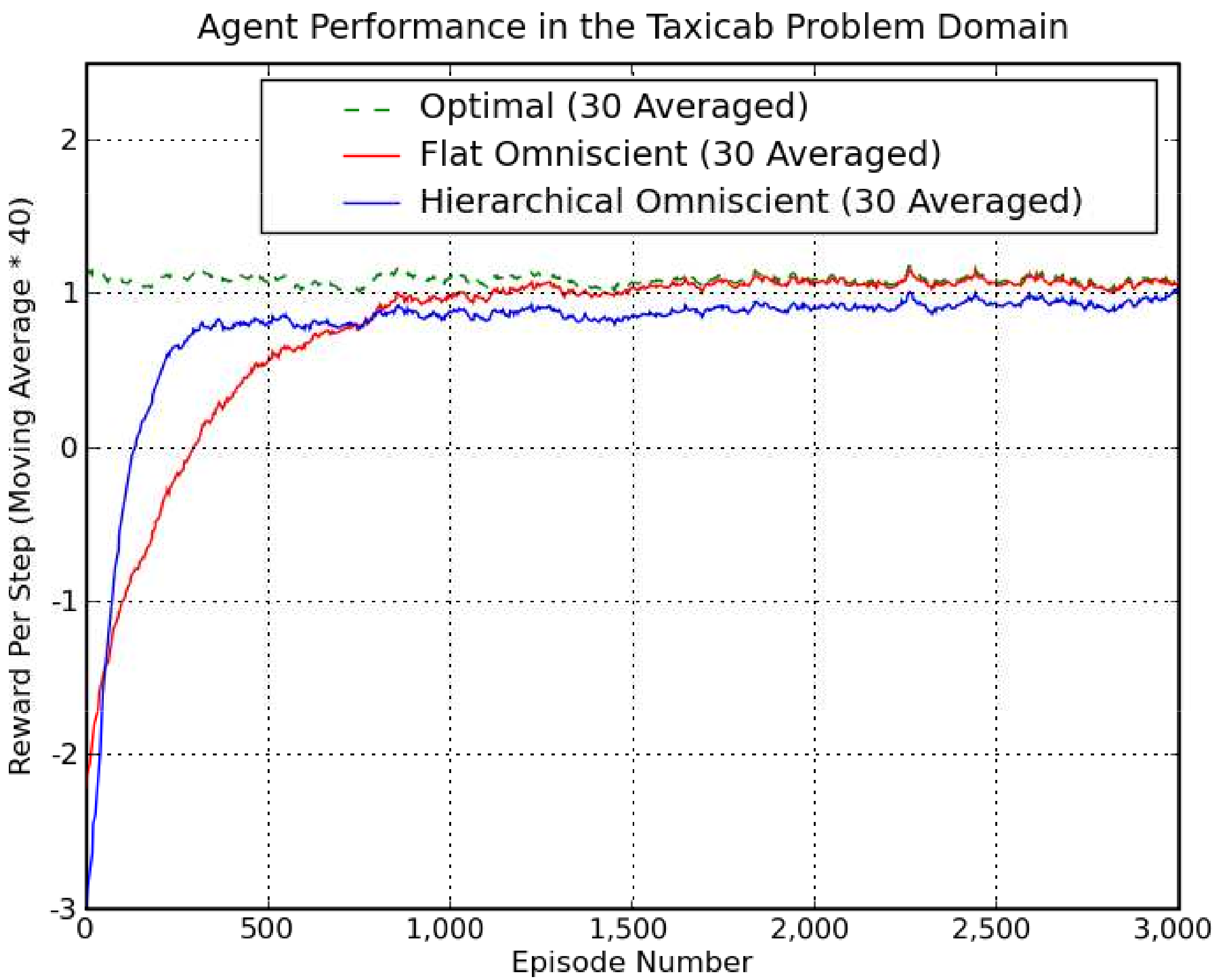
- Informed (Given The Passenger Source Color

  & The Passenger Destination Color)
  - Infinite Fuel
  - Finite Fuel ← Dietterich's Task
- Uninformed (Given Only Sensory Input)
  - Infinite Fuel
  - Finite Fuel

Four Agents

- Omniscient (Takes Advantage Of Given Source & Destination Color Information)
  - Flat
  - Hierarchical
- Uninformed (Must Search For The Passenger & Learn The Destination Upon Pickup)
  - Flat
  - Hierarchical

Informed-Infinite



Reward Per Step

(500 Extra Runs)

Optimal ≈ 1.10

Flat ≈ 1.09

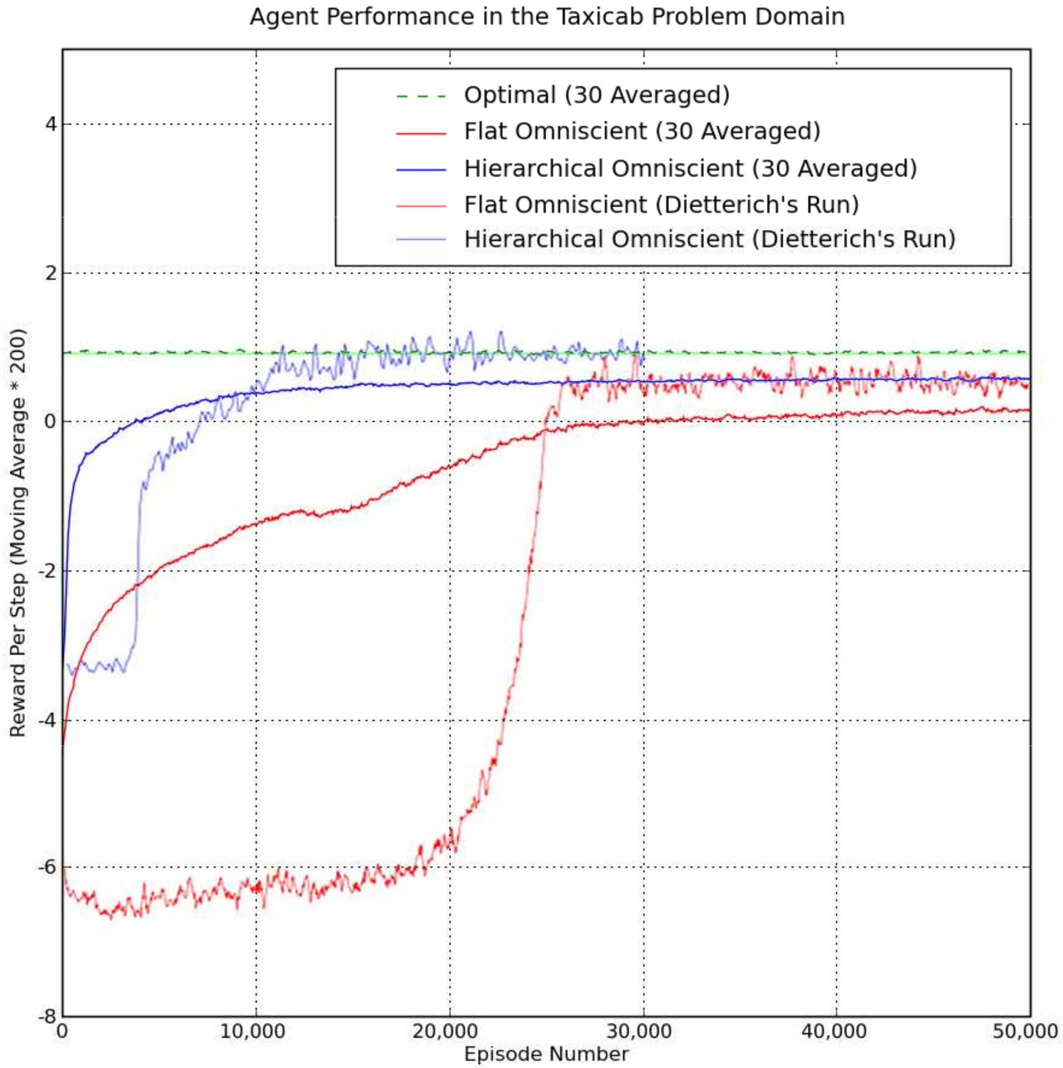
HRL ≈ 1.10

Optimal

Flat = 0 / 30

HRL = 29 / 30

Informed-Finite



Reward Per Step

(500 Extra Runs)

Optimal ≈ 0.93

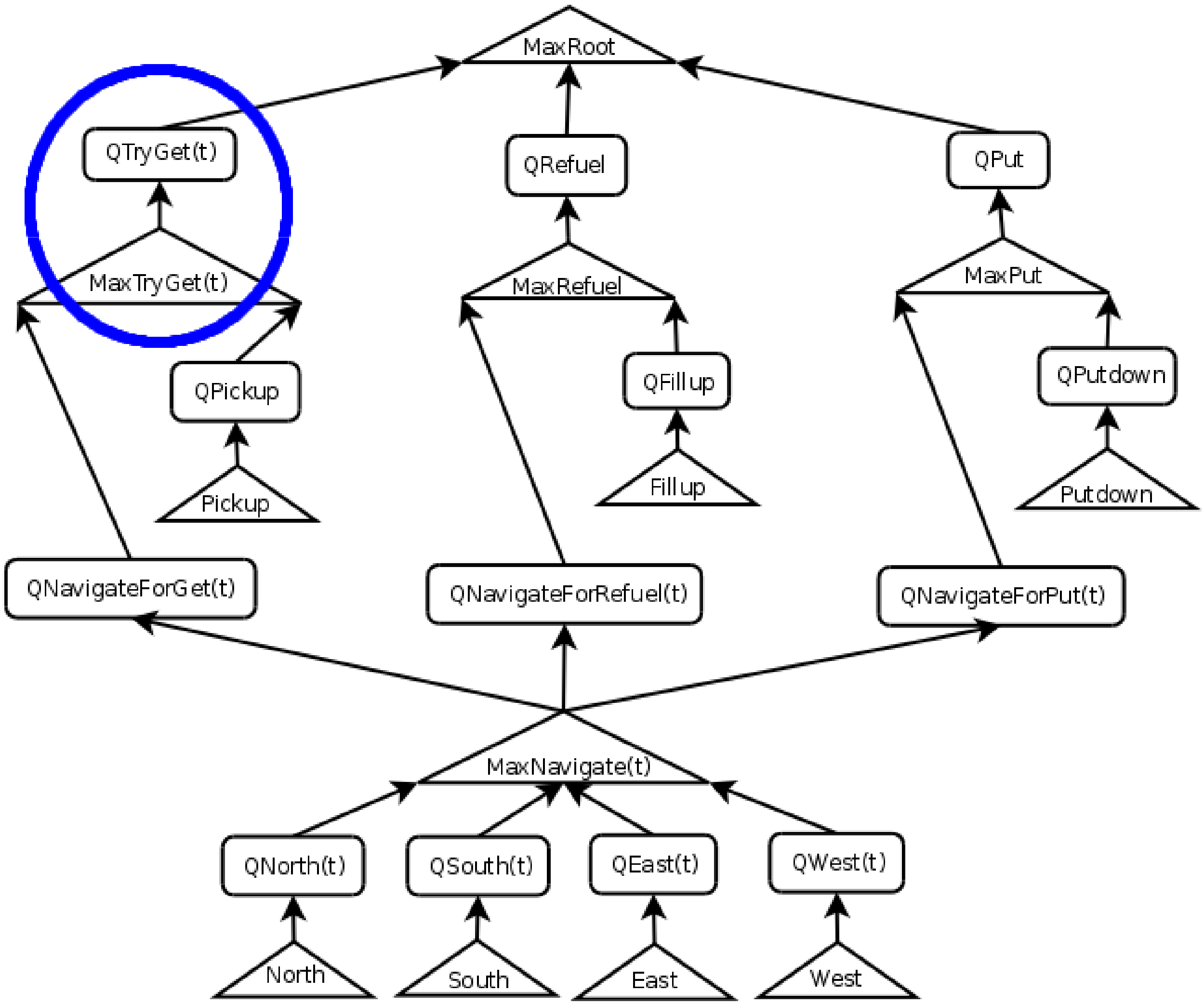
Flat ≈ 0.16

HRL ≈ 0.58

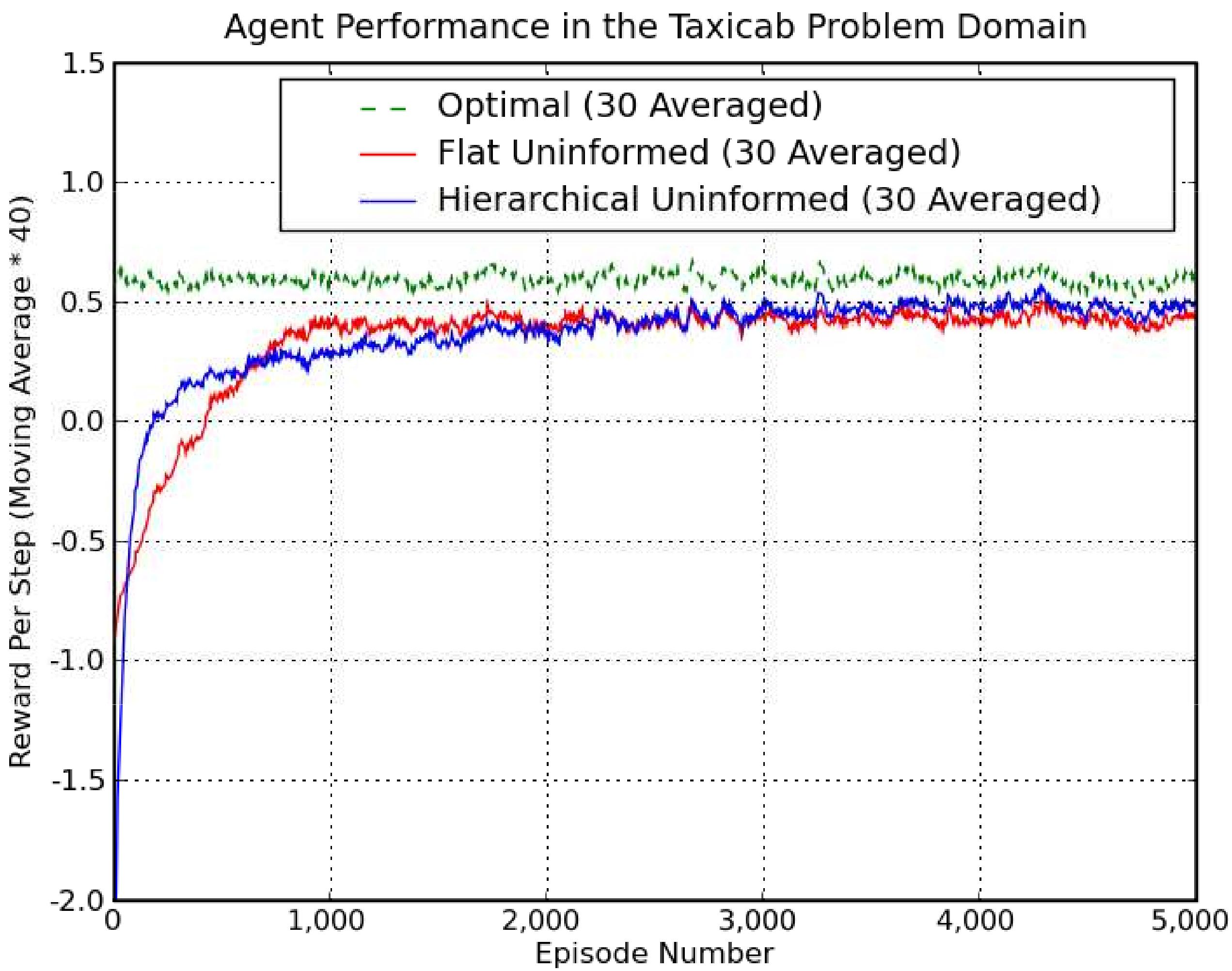
Optimal

None

Uninformed Hierarchy



Uninformed-Infinite



Reward Per Step

(500 Extra Runs)

Optimal ≈ 0.58

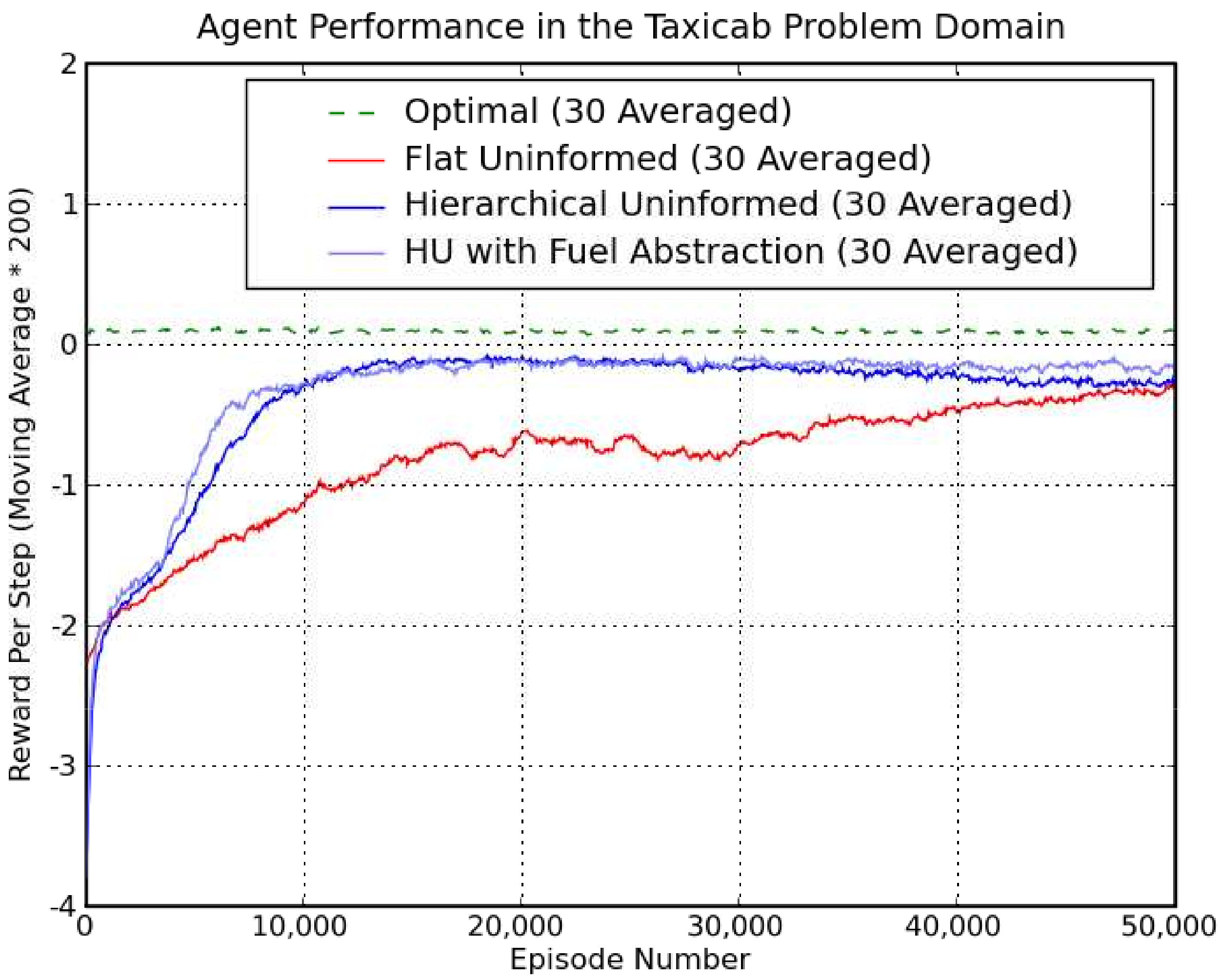
Flat ≈ 0.42

HRL ≈ 0.47

Optimal

None

Uninformed-Finite



Reward Per Step

(500 Extra Runs)

Optimal ≈ 0.10

Flat ≈ -0.29

HRL ≈ -0.27

HRL w/ FA ≈ -0.16

Fuel Abstraction =
0, 1, 2, 3, 4,
5-9, 10-13, 14

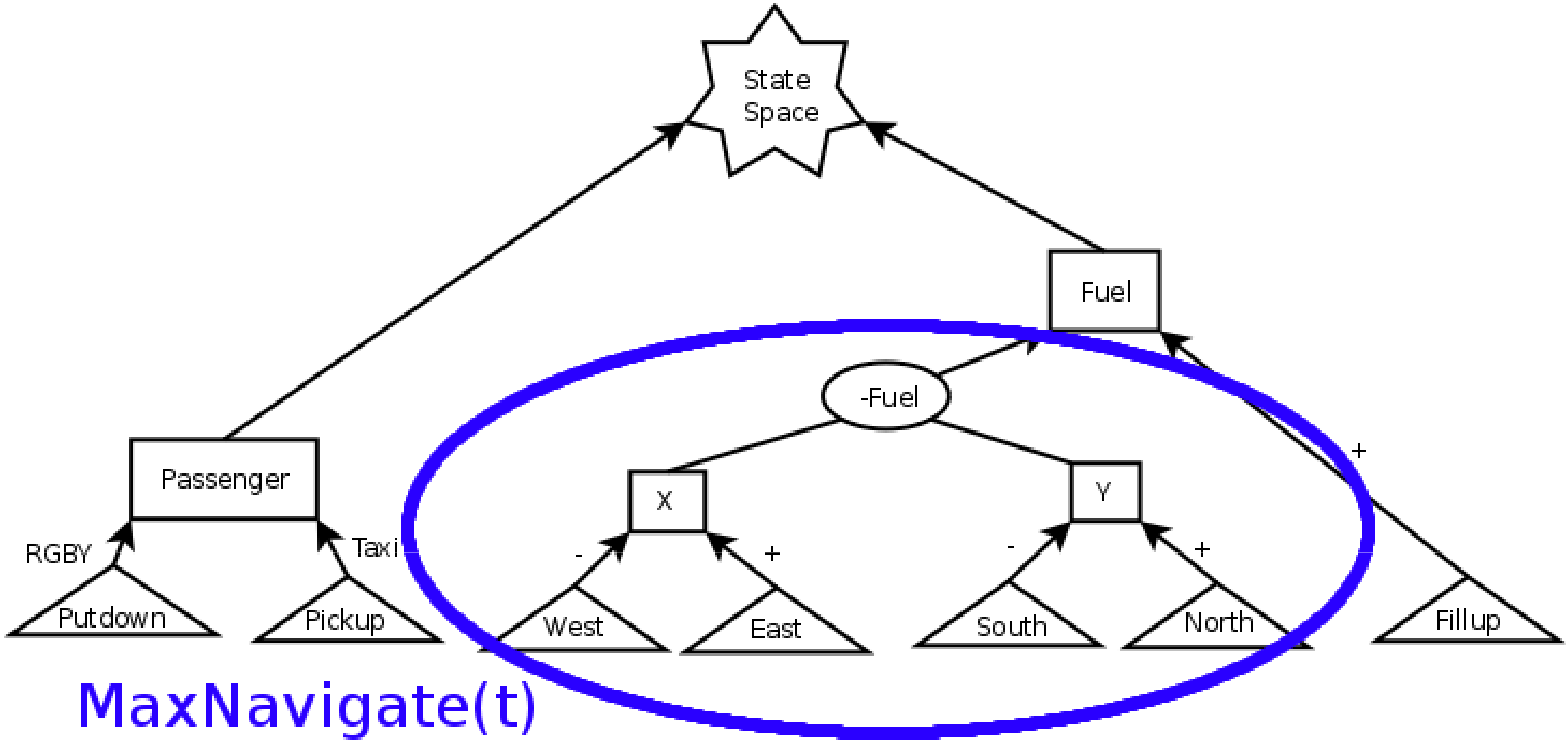
Observations

- MaxQ decomposition of the reward function is problematic when reward is undiscounted
  - Certain costs must affect multiple nodes
    - Additive property of the decomposition is violated
- Difficult to evaluate learning by direct analysis of reward rules in Soar
  - Certain types of decisions can visualized in an N-dimensional space (primitive motion decisions)
  - Others are more difficult to map to a visual (choice of next subtask)

Current / Future Work

- Automatic hierarchy generation
  - Factored State Representation
  - Given the result of each action from any given state, extrapolate hierarchical structure from trends in the changes in state variables
- Related work
  - Predictive State Representation
  - $\mathrm{DOOR}_{\mathrm{MAX}}$

Hierarchy Generation (In Progress)



Nuggets and Coal

- Nuggets
  - Soar-RL implementation of HRL is effective
  - SML allowed easy implementation of Dietterich's “one temperature per node” technique for HRL
- Coal
  - Verification of policy optimality is non-trivial
  - Uninformed-Hierarchical Agent unlearns the Finite-Fuel Task after 20,000 episodes