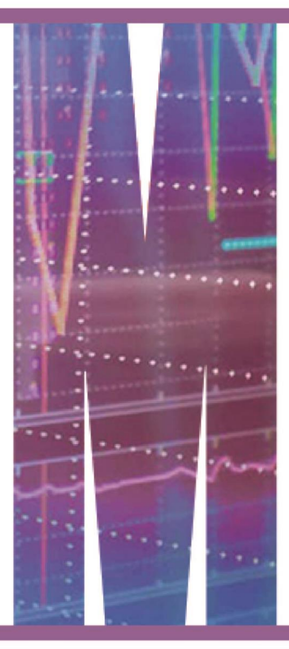
TAMING THE VALUE LANDSCAPE OF PREFERENCE LEARNING

Alexander Li

Supervised by:
Professor Michael G Burke



ECSE

INTRODUCTION

Reinforcement Learning (RL) can train robotics agents to complete physical tasks, by iteratively optimising a policy that takes in the immediate state and produces an optimal immediate action. Preference-based Reinforcement Learning (PbRL) allows the training process to be guided by human preference labels on state trajectories [1]. Probabilistic Temporal Ranking (PTR) is an instance of PbRL that compares single states, and automatically generates the preference labels assuming the later states in a demonstration are likely to have higher values than earlier states [2]. This approach is simple and scalable but has a problem: when the value predictions bleed out of state space barriers, the agent will learn to approach that counter-productive signal, costing training time and potentially never learning to reach the goal.

In this project, I show that attenuating the value landscape of PTR can reduce the counter-productive signal and allow the agent to learn to approach the goal along the correct trajectory.

METHOD

1. Construct a maze environment to emulate a state space with complex topology, along with limited observation $\boldsymbol{x}$ – the immediate coordinates.

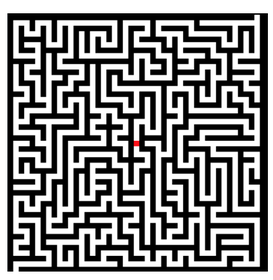


Figure 1. An example maze. The goal is marked with a red pixel.

2. Generate and record expert demonstrations of maze traversal. Create PTR dataset out of randomly sampled state pairs with preference labels.
3. Train a model with augmented input Fourier features using PTR. The model predicts a scalar $\boldsymbol{T(x)}$ representing the remaining progress to the goal and $\boldsymbol{V(x)}$ representing the value of the state:

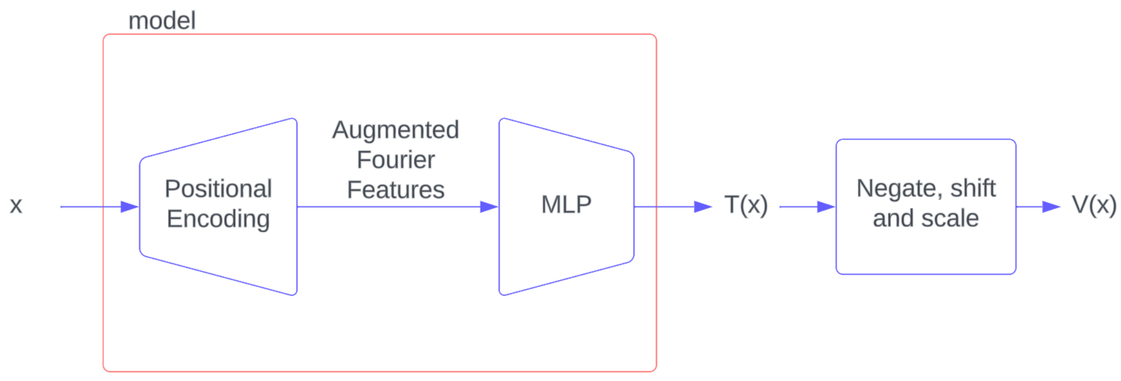


Figure 2. The model during inference. The less remaining progress, the higher the inferred value of that state.

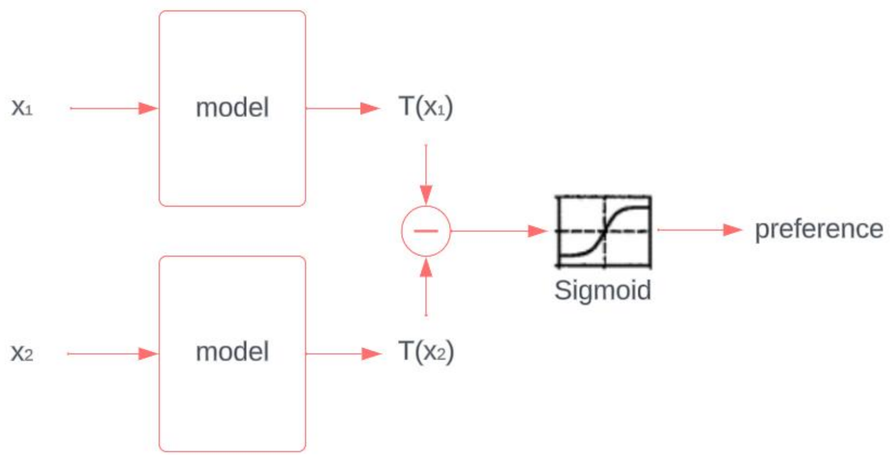


Figure 3. Forward pass of the PTR preference learning training step. A pair of states are passed into the same model whose outputs are compared.

4. Attenuate the value landscape using a Gaussian function with contraction factor $\boldsymbol{a}$:

$$V(x) \leftarrow V(x)e^{-aT(x)^2} \qquad (1)$$

5. Compare the performance of agents trained using various contraction factors of value attenuation.

RESULTS

- The model roughly captures the true value landscape (from the expert). However, there are regions where the incorrect ordering of the value landscape results in the wrong greedy action.
- The attenuation mechanism results in a natural contraction of the value signal towards the goal, while preserving the shape of the problem.
- The attenuated value map allows the agent to reach perfect performance **4000 epochs** before the sparse, unfiltered landscape.

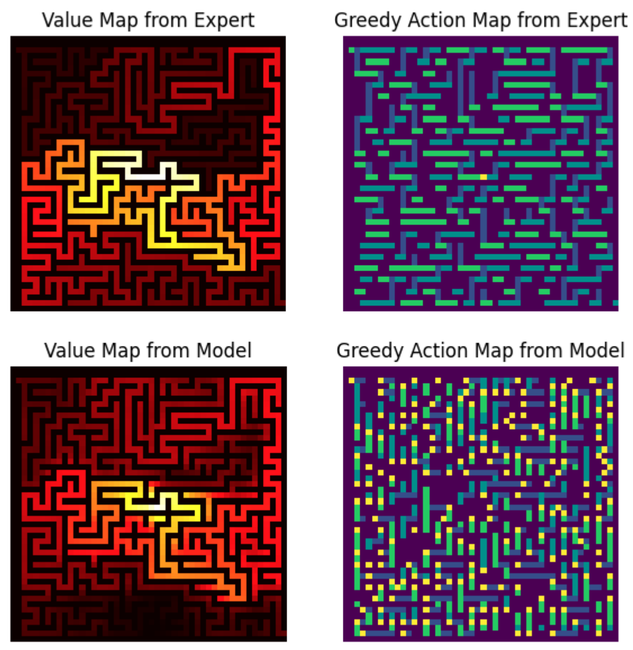


Figure 4. The true value map (from the expert) compared against the value map predicted by the model, and their associated greedy actions (dark purple = up, dark blue = down, dark green = left, light green = right, yellow = stay still)

Figure 5. Result of progressive attenuation of the value landscape, from left to right. Topology of the problem is preserved.

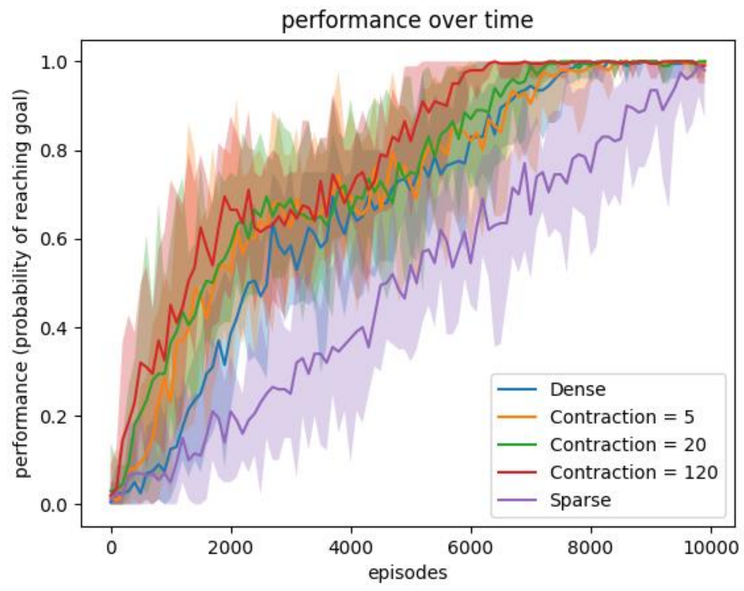


Figure 6. Each value map has 10 agents trained on them, and their performance plotted with 2-sigma uncertainty.

CONCLUSIONS

- The PTR output provides a natural attenuation mechanism that preserves the value landscape topology while amplifying the effect of reaching the goal on reinforcement.
- The attenuated value landscapes make agents learn faster.

REFERENCES

[1] Wirth, C., et al. "A survey of preference-based reinforcement learning methods," in Journal of Machine Learning Research, vol. 18, no. 136, pp. 1–46, 2017.

[2] Burke, M, et al, "Learning rewards for robotic ultrasound scanning using probabilistic temporal ranking," 2023.