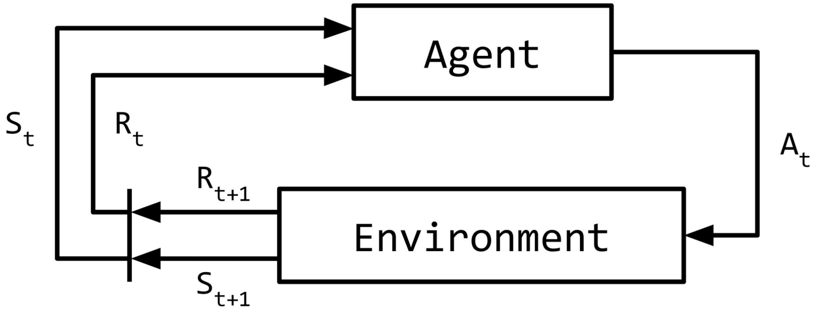
Maximizing Local Outlier Factors As an Objective for Exploration in Reinforcement Learning

Dylan R. Ashley

Outline

Reinforcement Learning

- Models an agent interacting with an environment
- Agent wants to maximize reward signal with later rewards being considered less valuable
- The agent doesn't know about it's environment beforehand and has to explore to learn about it

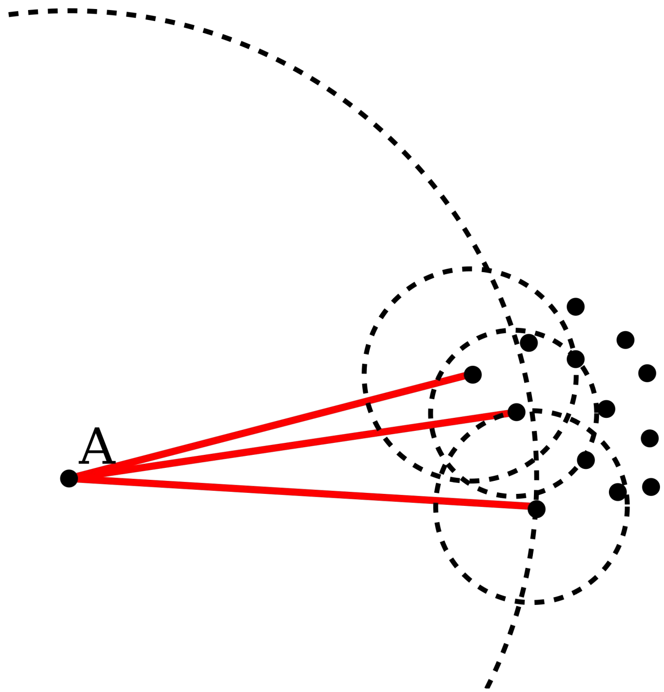


Local Outlier Factor (LOF)

- Compares the density around a point to the density around each neighbour in the point's *k*-NN

- LOF ≈ 1 means the point is probably in a cluster

- LOF >> 1 means the point is probably an outlier



Calculating the Local Outlier Factor

$$k\text{-distance}(p) = \max_{q \in k\mathrm{NN}(p)} d(p, q)$$

$$\text{reach-dist}_k(p, q) = \max\{\mathrm{d}(p, q), k\text{-distance}(q)\}$$

$$\mathrm{lrd}(p) = \frac{1}{\frac{1}{k}\sum_{q \in k\mathrm{NN}(p)} \text{reach-dist}_k(p, q)}$$

$$\mathrm{LOF}(p) = \frac{\frac{1}{k}\sum_{q \in k\mathrm{NN}(p)} \mathrm{lrd}(q)}{\mathrm{lrd}(p)}$$

Outline

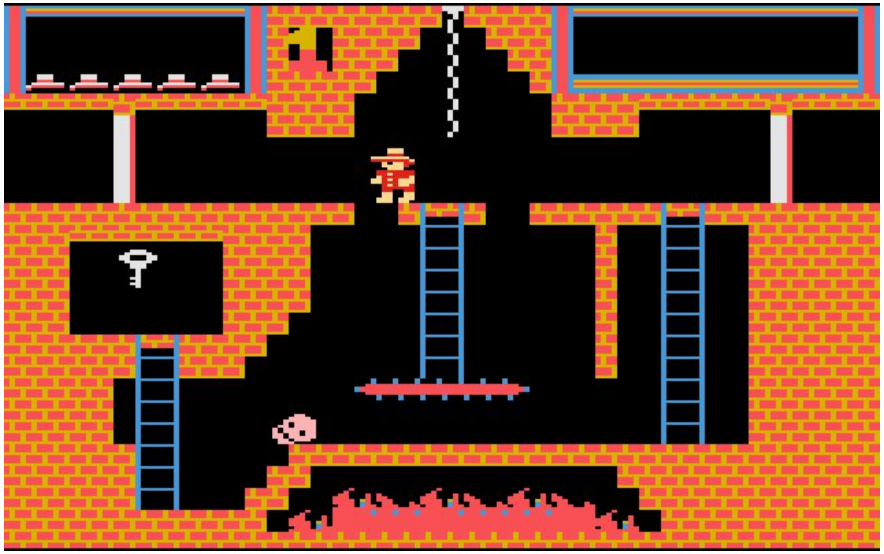
1. Background
    1.1. Reinforcement Learning
    1.2. Local Outlier Factor
2. Motivation
3. Basic Idea
4. Experimental Evaluation
    4.1. Setup
        4.1.1. Domain
        4.1.2. Algorithms
        4.1.3. Performance Metric
    4.2. Results
    4.3. Discussion

Motivation for Better Exploration

- Exploration is a major factor in the learning rate for an agent

- Without proper exploration the agent is liable to miss something important

- Naïve approaches often perform very badly



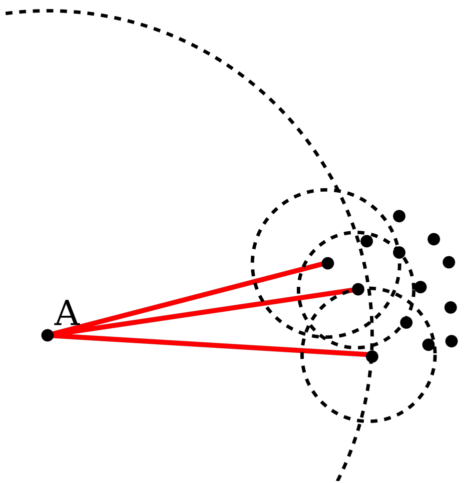
Outline

Using the LOF to Guide Exploration

- Reward an RL agent with the LOF of each state or transition it sees and a second agent can learn using the real rewards

- Intuitively this incentivises the agent to find rare states or transitions

- Can combine with random exploration to smooth out the exploration

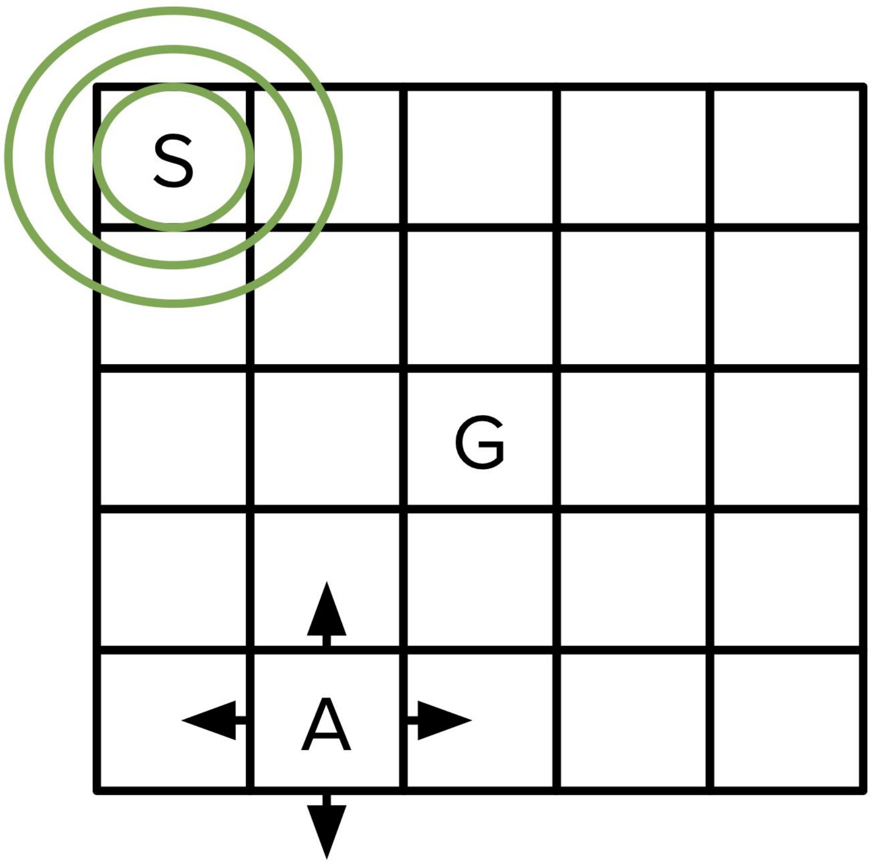


Outline

1. Background
    1.1. Reinforcement Learning
    1.2. Local Outlier Factor
2. Motivation
3. Basic Idea
4. Experimental Evaluation
    4.1. Setup
        4.1.1. Domain
        4.1.2. Algorithms
        4.1.3. Performance Metric
    4.2. Results
    4.3. Discussion

Domain

- 5x5 Gridworld with single start and terminal state
- Observations drawn from truncated Gaussian centered at current state scaled to cell size
- Empirically learnable with simple RL methods



Algorithms

- Q-learning with tile coding as function approximation

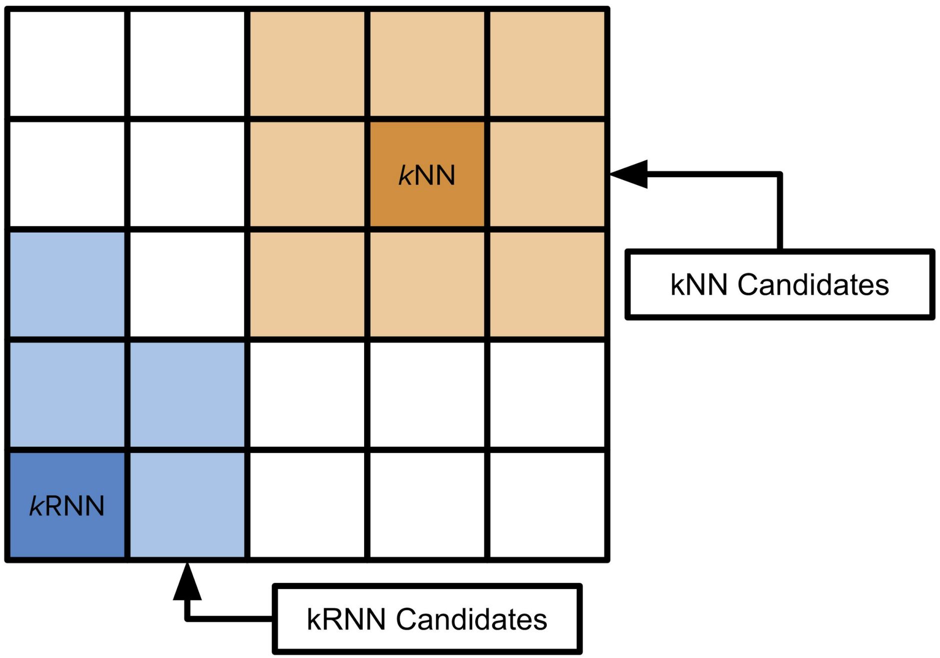
- Well known algorithm that learns the value of taking an action based on the state the agent is currently in

- Note that tile coding is an optimal strategy for guessing the true state in this domain

Initialize $Q(s, a)$ arbitrarily
Repeat (for each episode):
    Initialize s
    Repeat (for each step of episode):
        Choose a from s using policy derived from Q (e.g., ε-greedy)
        Take action a, observe r, s'
        $Q(s, a) \leftarrow Q(s, a) + \alpha\big[r + \gamma \max_{a'} Q(s', a') - Q(s, a)\big]$
        $s \leftarrow s'$;
    until s is terminal

Algorithms (continued)

- Use a grid for *k*-NN to prevent having to check all data points in *k*-NN search
- Increases overhead and slows early *k*-NN computation in exchange for faster later *k*-NN computation
- No change in worst case time complexity



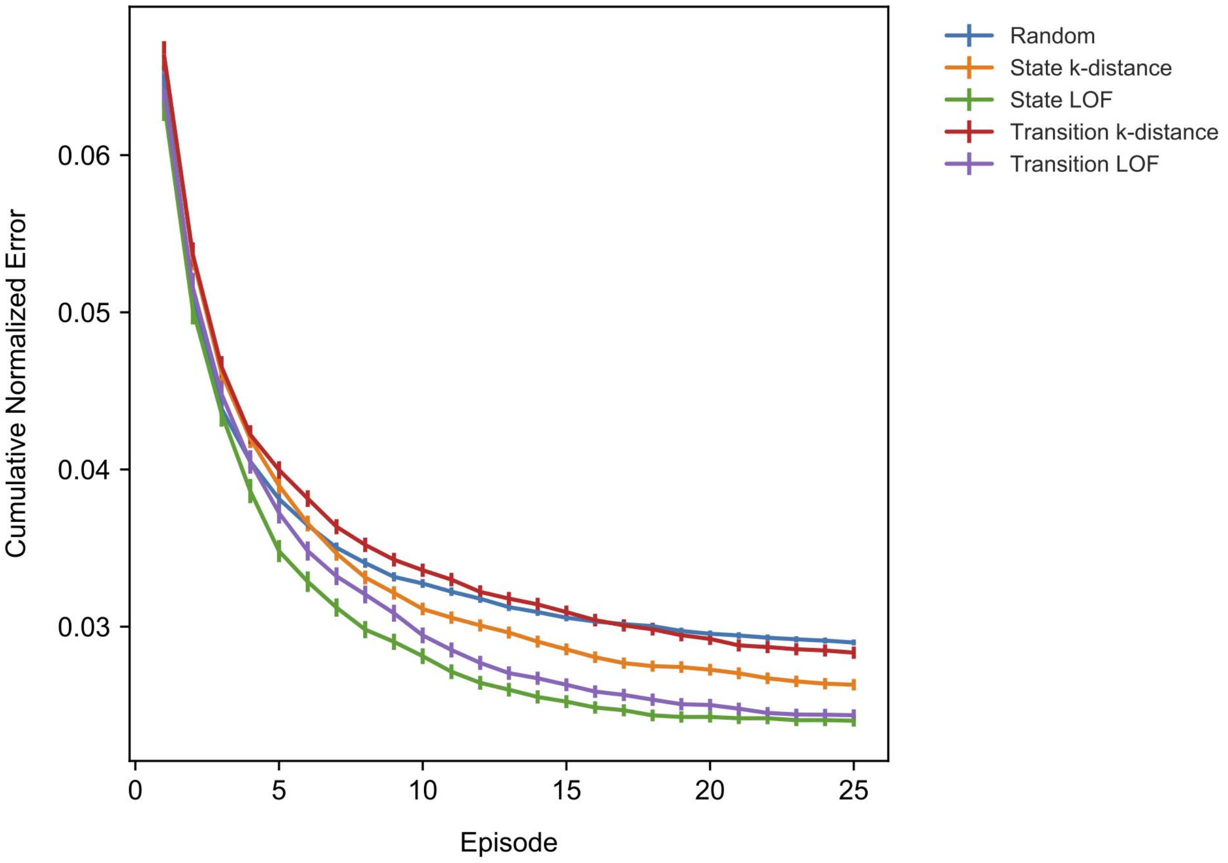
Performance Metric

$$\sigma = \sqrt{\frac{\sum_{i=1}^{|S|} \left(V_{S_i} - \overline{V} \right)^2}{|S| - 1}}$$

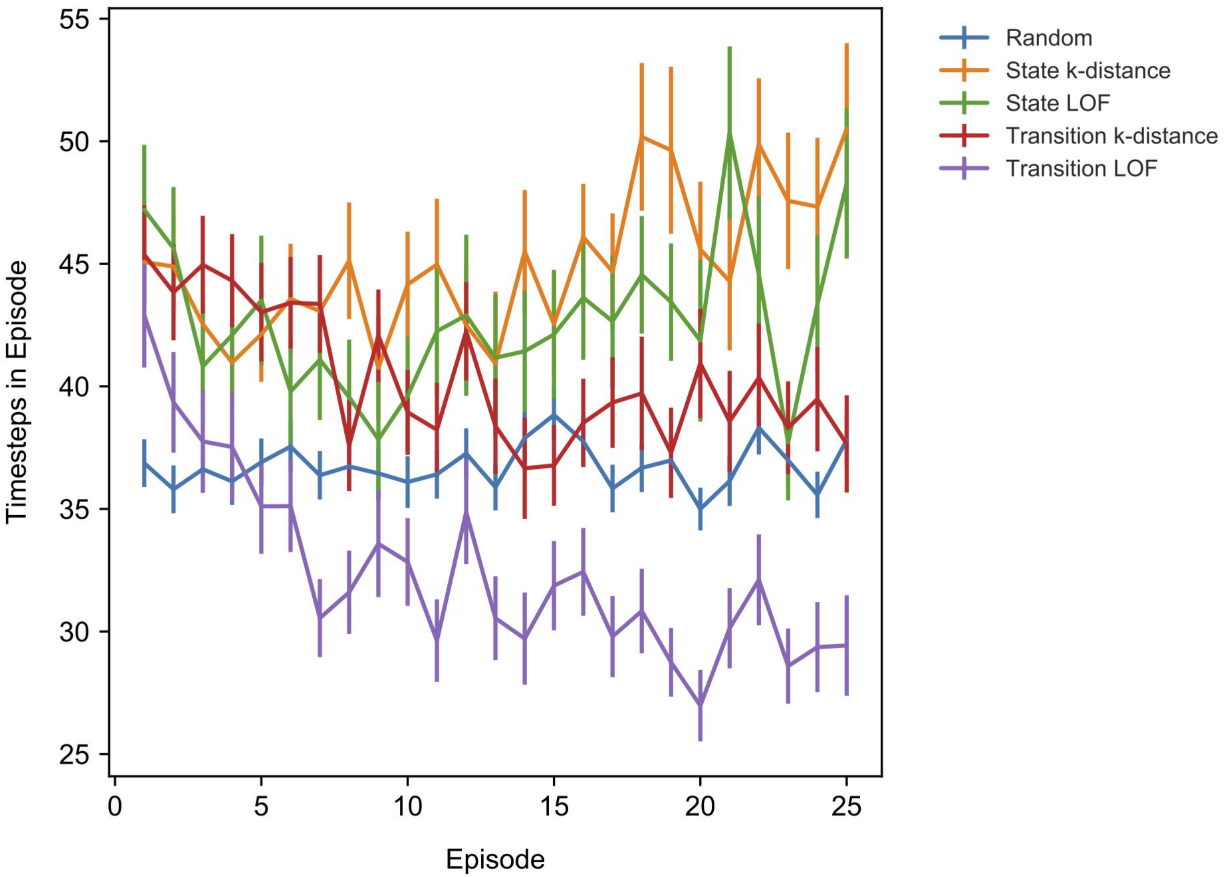
- Standard deviation of state visitations
- Minimizing means a more even exploration of the world
- Penalizes linear increase in distance to mean superlinearly

Outline

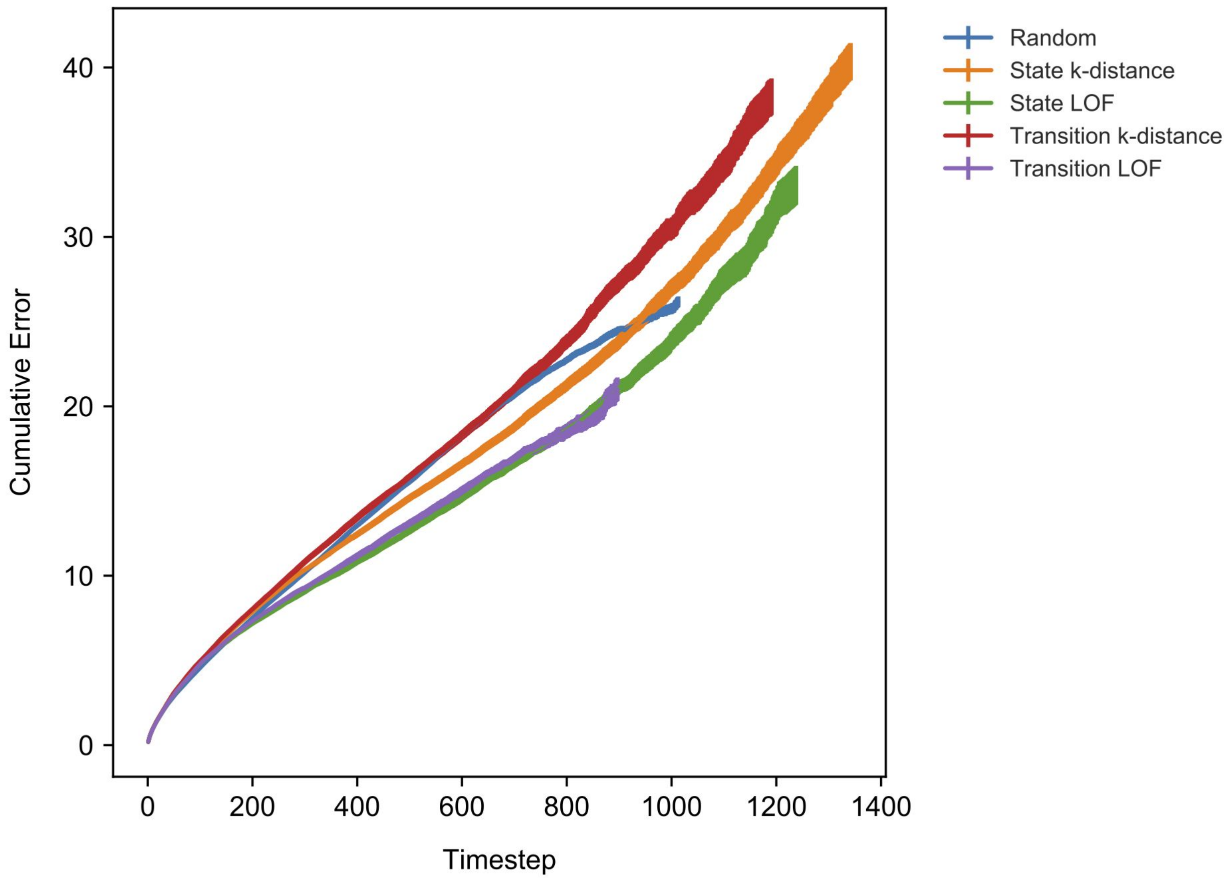
Experimental Results



Experimental Results (continued)



Experimental Results (continued)



Outline

Discussion

Pros

- Since it’s just a reward signal it can be used with any RL algorithm

- Natively handles a continuous space

- Can handle a changing environment by storing only the last *n* states or transitions

Cons

- Have to store states or transitions
    - If you only store the last n this is viable for the long term

- Doesn’t natively handle identical observations
    - Clipping values may be possible without incurring too severe a penalty

- *k*-NN is quite expensive
    - There are cheaper approaches that haven’t been considered for this project because of complexity

References

- https://upload.wikimedia.org/wikipedia/commons/thumb/4/4e/LOF-idea.svg/2000px-LOF-idea.svg.png
- https://atariage.com/5200/screenshots/s_MontezumasRevenge_2.png
- http://img.taste.com.au/BcemwIdD/taste/2016/11/chocolate-celebration-cake-85607-1.jpeg
- https://qph.ec.quoracdn.net/main-qimg-581fc72f6fa620741c9119688fafd5c6