Scalable Continuous Reinforcement Learning with Kullback Leibler Divergence Policy Chain

Christopher Doyle, Undergraduate

Reinforcement Learning (RL) is an area of machine learning that involves an agent optimising its traversal through a Markov Decision Process (MDP) in such a way that rewards are maximised. Today's standard RL value function methods (Q-Learning), policy methods (Policy Gradient), and compound methods (Actor-Critic) primarily focus on training an agent to learn one task optimally. The problem of catastrophic forgetting can arise after optimisation when the policy remains constant but either 1) the rewards are changed or 2) the task and environment are changed. This article addresses the former issue (1) and attempts to improve upon existing 'continuous' RL techniques by exploring a scalable approach adopting techniques used in Information Theory. A dependent chain of policies is forged with 'memories' from a Kullback Leibler Divergence Loss Function of existing policies. In summary, this article explores extracting the useful 'memories' from previously learned policies and utilising these skills from any new task's starting point. I want to make my approach more scalable than the dictionary approach of storing all learned policies by creating what I call a Kullback-Leibler Divergence Policy Chain.

1 Introduction

Catastrophic Forgetting is the phenomenon that occurs when a mathematical model's accuracy will suddenly fall dramatically and result is an agent 'forgetting' how to perform a task. In the context of RL, this event often occurs when an agent is asked to perform a new task, B, after perfecting task A. Agents can successfully learn how to optimise performance in B after learning A but if they are asked to once again perform task A, then the agent's policy can diverge and the agent can become unable to learn either A or B.

The exact problem that I want to explore a solution to is when a task's nature is changed via changing the reward system and an agent is forced to adapt and learn to re-optimise the environment with a converging policy. The environment I will focus on is the GridWorld setting and the agent will experience a reward of -1 for every step it takes until the goal and a reward of -50 for going out of bounds. The problem is as follows:

1. Set a goal location
2. Choose a suitable starting policy
3. Optimise the policy
4. Repeat

Step 2. is what this article will primarily focus on. We want to 'choose' a suitable policy. Dictionary approaches such as the 'Composition of value functions' concern the collection of optimised value functions and selecting the most suitable with which to begin a new task. I want to address the scalability issues associated with this method in that learned policies or value functions take up space. This article's method requires at most three polices be stored at any one time. The continuous reinforcement learning (CRL) technique of Variational Continual Learning is a variational inference approach (that uses Bayesian neural networks (MacKay, 1992; Neal, 1995). VCL sets the posterior at the end of training a task to be the prior when beginning training for the next task.) My approach of a Kullback-Leibler policy chain is similar in the regard that the posterior distribution at the end of one task contributes to the prior of the next task but it is not the only task to contribute. This is made possible by the properties of the Kullback-Leibler divergence. My proposed method is as follows:

Set-Up Block (first three polices π_1, π_2, π_3)

1. Choose a random goal position in the grid.

2. Assume prior policy $\pi_1 \sim U(0, L)$.

3. Optimise the posterior π_1 for task 1, then save a copy of π_1^*.

4. Repeat from (1.) for task 2.

5. Choose a random goal position in the grid (i.e. task 3).

6. Assume prior policy

$\pi_3 \sim \min_{\pi_\theta} \mathcal{L}\left(\theta|\pi_A = \pi_1^*, \pi_B = \pi_2^*\right)$.

7. Optimise the posterior π_3 for task 3, then save a copy of π_3^*.

Main Block (all other tasks to be learned)

1. Change the goal (i.e. create task $i, i_{initial} = 4$).

2. Assume prior distribution
   $\pi_i \sim \min_{\pi_\theta} \mathcal{L}\left(\theta|\pi_A = \pi_{i-1}, \pi_B = \pi_{i-1}^*\right)$.

3. i.e. for task 4, prior
   $\pi_4 \sim \min_{\pi_\theta} \mathcal{L}\left(\theta|\pi_A = \pi_3, \pi_B = \pi_3^*\right)$.

The remaining requirement of this method is then to define the loss function $\mathcal{L}\left(\theta|\pi_A, \pi_B\right)$.

My intention for this loss function's purpose is to ensure that any new policy $\pi_{i,new}$ will primarily remember the useful elements of previously learned policies, and will converge to a uniform policy where learned polices greatly differ. This will prove useful in that an agent should 'remember' not to go out of bounds in an area a previous policy has explored. A problematic scenario that can arise is that old 'memories' might pull our agent away from where it needs to go if our goal has not been near that area in the past. This problem should become negligible as the number of tasks we learn becomes large (see background).

We define $\mathcal{L}\left(\theta|\pi_A, \pi_B\right)$ as

$$\mathcal{L}(\theta) = \alpha \mathcal{D}_{KL}\left(\pi(\theta) \mid\mid \pi_A\right) + (1-\alpha)\mathcal{D}_{KL}\left(\pi(\theta) \mid\mid \pi_B\right)$$

and any new prior policy $\pi_{i,new}$ as

$$\pi_i \sim \min_{\pi_\theta} \mathcal{L}\left(\theta|\pi_A = \pi_{i-1}, \pi_B = \pi_{i-1}^*\right) \quad , i \geq 3.$$

In this was we forge a dependent policy 'chain'.

2 Background

2.1 Reinforcement Learning

2.2 Policy Gradients

2.3 Kullback-Leibler Divergence

Aliquam elementum nulla at arcu finibus aliquet. Praesent congue ultrices nisl pretium posuere. Nunc vel nulla hendrerit, ultrices justo ut, ultrices sapien. Duis ut arcu at nunc pellentesque consectetur. Vestibulum eget nisl porta, ultricies orci eget, efficitur tellus. Maecenas rhoncus purus vel mauris tincidunt, et euismod nibh viverra. Mauris ultrices tellus quis ante lobortis gravida. Duis vulputate viverra erat, eu sollicitudin dui. Proin a iaculis massa. Nam at turpis in sem malesuada rhoncus. Aenean tempor risus dui, et ultrices nulla rutrum ut. Nam commodo fermentum purus, eget mattis odio fringilla at. Etiam congue et ipsum sed feugiat. Morbi euismod ut purus et tempus. Etiam est ligula, aliquam eget porttitor ut, auctor in risus. Curabitur at urna id dui lobortis pellentesque.

$$A = \begin{bmatrix} A_{11} & A_{21} \\ A_{21} & A_{22} \end{bmatrix} \tag{1}$$

Donec nec nibh sagittis, finibus mauris quis, laoreet augue. Maecenas aliquam sem nunc, vel semper urna hendrerit nec. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Maecenas pellentesque dolor lacus, sit amet pretium felis vestibulum finibus. Duis tincidunt sapien faucibus nisi vehicula tincidunt. Donec euismod suscipit ligula a tempor. Aenean a nulla sit amet magna ullamcorper condimentum. Fusce eu velit vitae libero varius condimentum at sed dui.

In non leo tincidunt, tristique orci eu, suscipit ex. Fusce non lectus ut dolor tincidunt fermentum. Donec dictum mauris magna, ut dictum nisl finibus quis. Nulla elementum ipsum ut lectus sodales finibus. Nulla ac malesuada magna. Etiam arcu dolor, luctus eget elit a, volutpat vulputate mi. Donec elementum tellus libero, ut ornare orci dignissim lacinia. Nullam iaculis vehicula sem, at tempor tellus. Praesent eu nisi a elit viverra lobortis. Nullam eu metus et justo molestie posuere vitae imperdiet erat. Praesent at gravida dui. Vivamus mauris odio, efficitur eget lacus quis, mattis tristique risus. Mauris quis metus sed risus lobortis sollicitudin vitae vitae quam. Morbi leo turpis, aliquam at nunc sit amet, ultricies dictum lorem. Nam et fringilla elit. Vestibulum auctor, turpis ut facilisis tempor, arcu nibh tincidunt libero, quis blandit leo turpis a urna.

2.4 Subsection

Nam ante risus, tempor nec lacus ac, congue pretium dui. Donec a nisl est. Integer accumsan mauris eu ex venenatis mollis. Aliquam sit amet ipsum laoreet, mollis sem sit amet, pellentesque quam. Aenean auctor diam eget erat venenatis laoreet. In ipsum felis, tristique eu efficitur at, maximus ac urna. Aenean pulvinar eu lorem eget suscipit. Aliquam et lorem erat. Nam fringilla ante risus, eget convallis nunc pellentesque non. Donec ipsum nisl, consectetur in magna eu, hendrerit pulvinar orci. Mauris porta convallis neque, non viverra urna pulvinar ac. Cras non condimentum lectus. Aliquam odio leo, aliquet vitae tellus nec, imperdiet lacinia turpis. Nam ac lectus imperdiet, luctus nibh a, feugiat urna.

- First item in a list
- Second item in a list
- Third item in a list

Nunc egestas quis leo sed efficitur. Donec placerat, dui vel bibendum bibendum, tortor ligula auctor elit, aliquet pulvinar leo ante nec tellus. Praesent at vulputate libero, sit amet elementum magna. Pellentesque sodales odio eu ex interdum molestie. Suspendisse lacinia, augue quis interdum posuere, dolor ipsum euismod turpis, sed viverra nibh velit eget dolor. Curabitur consectetur tempus lacus, sit amet luctus mauris interdum vel. Curabitur vehicula convallis felis, eget mattis justo rhoncus eget. Pellentesque et semper lectus.

First This is the first item
Last This is the last item

Donec nec nibh sagittis, finibus mauris quis, laoreet augue. Maecenas aliquam sem nunc, vel semper urna hendrerit nec. Pellentesque habitant morbi tristique senectus et netus et malesuada fames ac turpis egestas. Maecenas pellentesque dolor lacus, sit amet pretium felis vestibulum finibus. Duis tincidunt sapien faucibus nisi vehicula tincidunt. Donec euismod suscipit ligula a tempor. Aenean a nulla sit amet magna ullamcorper condimentum. Fusce eu velit vitae libero varius condimentum at sed dui.

2.5 Subsection

In hac habitasse platea dictumst. Etiam ac tortor fermentum, ultrices libero gravida, blandit metus. Vivamus sed convallis felis. Cras vel tortor sollicitudin, vestibulum nisi at, pretium justo. Curabitur placerat elit nunc, sed luctus ipsum auctor a. Nulla feugiat quam venenatis nulla imperdiet vulputate non faucibus lorem. Curabitur mollis diam non leo ullamcorper lacinia.

Morbi iaculis posuere arcu, ut scelerisque sem. Class aptent taciti sociosqu ad litora torquent per conubia nostra, per inceptos himenaeos. Mauris placerat urna id enim aliquet, non consequat leo imperdiet. Phasellus at nibh ut tortor hendrerit accumsan. Phasellus sollicitudin luctus sapien, feugiat facilisis risus consectetur eleifend. In quis luctus turpis. Nulla sed tellus libero. Pellentesque metus tortor, convallis at tellus quis, accumsan faucibus nulla. Fusce auctor eleifend volutpat. Maecenas vel faucibus enim. Donec venenatis congue congue. Integer sit amet quam ac est aliquam aliquet. Ut commodo justo sit amet convallis scelerisque.

1. First numbered item in a list
2. Second numbered item in a list
3. Third numbered item in a list

Aliquam elementum nulla at arcu finibus aliquet. Praesent congue ultrices nisl pretium posuere. Nunc vel nulla hendrerit, ultrices justo ut, ultrices sapien. Duis ut arcu at nunc pellentesque consectetur. Vestibulum eget nisl porta, ultricies orci eget, efficitur tellus. Maecenas rhoncus purus vel mauris tincidunt, et euismod nibh viverra. Mauris ultrices tellus quis ante lobortis gravida. Duis vulputate viverra erat, eu sollicitudin dui. Proin a iaculis massa. Nam at turpis in sem malesuada rhoncus. Aenean tempor risus dui, et ultrices nulla rutrum ut. Nam commodo fermentum purus, eget mattis odio fringilla at. Etiam congue et ipsum sed feugiat. Morbi euismod ut purus et tempus. Etiam est ligula, aliquam eget porttitor ut, auctor in risus. Curabitur at urna id dui lobortis pellentesque.

Table 1: *Example table*

Name		
First Name	Last Name	Grade
John	Doe	7.5
Richard	Miles	5

3 Section

In hac habitasse platea dictumst. Vivamus eu finibus leo. Donec malesuada dui non sagittis auctor. Aenean condimentum eros metus. Nunc tempus id velit ut tempus. Quisque fermentum, nisl sit amet consectetur ornare, nunc leo luctus leo, vitae mattis odio augue id libero. Mauris quis lectus at ante scelerisque sollicitudin in eu nisi. Nulla elit lacus, ultricies eu erat congue, venenatis semper turpis. Ut nec venenatis velit. Mauris lacinia diam diam, ac egestas neque sodales sed. Curabitur eu diam nulla. Duis nec turpis finibus, commodo diam sed, bibendum erat. Nunc in velit ullamcorper, posuere libero a, mollis mauris. Nulla vehicula quam id tortor ornare blandit. Aenean maximus tempor orci ultrices placerat. Aenean condimentum magna vulputate erat mattis feugiat.

Quisque lacinia, purus id mattis gravida, sem enim fringilla erat, non dapibus est tellus pellentesque velit. Vivamus pretium sem quis leo placerat, at dignissim ex iaculis. Donec neque tortor, pharetra quis vestibulum id, tempus scelerisque mi. Cras in mattis est. Integer nec lorem rutrum, semper ligula bibendum, iaculis neque. Sed in nunc placerat, viverra dui in, fringilla sem. Sed quis rutrum magna, vitae pellentesque eros.

Praesent maximus mauris vitae nisl pulvinar, at tristique tortor aliquam. Etiam sit amet nunc in nulla vulputate sollicitudin. Aliquam erat volutpat. Praesent pharetra gravida cursus. Quisque vulputate lacus nunc. Integer orci ex, porttitor quis sapien id, eleifend gravida mi. Etiam efficitur justo eget nulla congue mattis. Duis commodo vel arcu a pretium. Aenean eleifend viverra nisl, nec ornare lacus rutrum in.

Vivamus pulvinar ac eros eu pellentesque. Duis nibh felis, sagittis sed lacus at, sagittis mattis nisi. Fusce ante dui, tincidunt in scelerisque ut, sagittis at magna. Fusce tincidunt felis et odio tincidunt imperdiet. Cras ut facilisis nisl. Aliquam vitae consequat metus, eget gravida augue. In imperdiet justo quis nulla venenatis accumsan. Aliquam aliquet consectetur tortor, at sollicitudin sapien porta sed. Donec efficitur mauris id rhoncus volutpat. Vestibulum ante ipsum primis in faucibus orci luctus et ultrices posuere cubilia Curae; Sed bibendum purus dapibus tincidunt euismod. Nullam malesuada ultrices lacus, ut tincidunt dolor. Etiam imperdiet quam eget elit tincidunt scelerisque. Curabitur ut ullamcorper dui. Cras gravida porta leo, ut lobortis nisl venenatis pulvinar. Proin non semper nulla.

Praesent pretium nisl purus, id mollis nibh efficitur sed. Sed sit amet urna leo. Nulla sed imperdiet sem. Donec ut diam tristique, faucibus ligula vel, varius est. In ipsum ligula, elementum vitae velit ac, viverra tincidunt enim. Phasellus gravida diam id nisl interdum maximus. Ut semper, tortor vitae congue pharetra, justo odio commodo urna, vel tempus libero ex et risus. Vivamus commodo felis non venenatis rutrum. Sed pulvinar scelerisque augue in porta. Sed maximus libero nec tellus malesuada elementum. Proin non augue posuere, pellentesque felis viverra, varius urna. Lorem ipsum dolor sit amet, consectetur adipiscing elit. Donec dignissim urna eget diam dictum, eget facilisis libero pulvinar.

Aliquam ex tellus, hendrerit sed odio sit amet, facilisis elementum enim. Suspendisse potenti. Integer molestie ac augue sit amet fermentum. Vivamus ultrices ante nulla, vitae venenatis ipsum ullamcorper sed. Phasellus gravida felis sapien, ac porta purus pharetra quis. Sed eget augue tellus. Nam vitae hendrerit arcu, id iaculis ipsum. Pellentesque sed magna tortor.

In ac tempus diam. Sed nec lobortis massa, suscipit accumsan justo. Quisque porttitor, ligula a semper euismod, urna diam dictum sem, sed maximus risus purus sit amet felis. Fusce elementum maximus nisi a mattis. Nulla vitae elit erat. Integer sit amet commodo risus, eget elementum nulla. Donec ultricies erat sit amet sem commodo iaculis. Donec euismod volutpat lacus, ut tempor est lacinia a. Vivamus auctor condimentum tincidunt. Praesent sed finibus urna. Sed pellentesque blandit magna et rhoncus.

Integer vel turpis nec tellus sodales malesuada a vel odio. Fusce et lectus eu nibh rhoncus tempus vel nec elit. Suspendisse commodo orci velit, lacinia dictum odio accumsan et. Vivamus libero dui, elementum vel nibh non, fermentum venenatis risus. Aliquam sed sapien ac orci sodales tempus a eget dui. Morbi non dictum tortor, quis tincidunt nibh. Proin ut tincidunt odio.

Pellentesque ac nisi dolor. Pellentesque maximus est arcu, eu scelerisque est rutrum vitae. Mauris ullamcorper vulputate vehicula. Praesent fermentum leo ac velit accumsan consectetur. Aliquam eleifend ex eros, ut lacinia tellus volutpat non. Pellentesque sit amet cursus diam. Maecenas elementum mattis est, in tincidunt ex pretium ac. Integer ultrices nunc rutrum, pretium sapien vitae, lobortis velit.