An analysis of the Pheromone Q-Learning algorithm

Ndedi Monekosso and Paolo Remagnino

Digital Imaging Research Centre
School of Computing and Information Systems
Kingston University, United Kingdom
{n.monekosso, p.remagnino}@kingston.ac.uk

Abstract. The Phe-Q machine learning technique, a modified Q-learning technique, was developed to enable co-operating agents to communicate in learning to solve a problem. The Phe-Q learning technique combines Q-learning with synthetic pheromone to improve on the speed of convergence. The Phe-Q update equation includes a belief factor that reflects the confidence the agent has in the pheromone (the communication) deposited in the environment by other agents. With the Phe-Q update equation, speed of convergence towards an optimal solution depends on a number parameters including the number of agents solving a problem, the amount of pheromone deposited, and the evaporation rate. In this paper, work carried out to optimise speed of learning with the Phe-Q technique is described. The objective was to to optimise Phe-Q learning with respect to pheromone deposition rates, evaporation rates.

Track

Paper

Conference Topics

Conference topics : Machine Learning, Multi-Agent Systems and Distributed AI

An analysis of the Pheromone Q-Learning algorithm

Abstract. The Phe-Q machine learning technique, a modified Q-learning technique, was developed to enable co-operating agents to communicate in learning to solve a problem. The Phe-Q learning technique combines Q-learning with synthetic pheromone to improve on the speed of convergence. The Phe-Q update equation includes a belief factor that reflects the confidence the agent has in the pheromone (the communication) deposited in the environment by other agents. With the Phe-Q update equation, speed of convergence towards an optimal solution depends on a number parameters including the number of agents solving a problem, the amount of pheromone deposited, and the evaporation rate. In this paper, work carried out to optimise speed of learning with the Phe-Q technique is described. The objective was to to optimise Phe-Q learning with respect to pheromone deposition rates, evaporation rates.

1 Introduction

In situations where building a partial or complete model of a complex problem and/or environment is impractical or impossible, learning to solve the problem may be the only solution. In addition, it is often convenient when solving complex or large problems, to break down the problem into its component parts, and each part dealt with separately. This technique is often seen in nature and one such example is the ant foraging behaviour. The ant colony exhibits a collective problem solving ability [4, 9]. Complex behaviours emerge from the interaction of the relatively simple behaviour of individual ants. A characteristic that multi-agent systems seek to reproduce. The ant colony exhibits among other features, co-operation and co-ordination, and communicate implicitly by depositing pheromone chemicals. The ant foraging for food will deposit a trail of pheromone. The problem for the ants is to find the shortest path between the nest and the food source whilst minimising energy. The Phe-Q algorithm was inspired by the search strategies of foraging ants, combining reinforcement learning [3, 19] with synthetic pheromone. Reinforcement learning has been applied with some success to classical A.I. problems. With this technique, the agent learns by trial and error. Phe-Q algorithm is described in [16]. The performance of the Phe-Q agent is dependent on a few factors namely pheromone deposition and diffusion rates, pheromone evaporation rate. Furthermore, in the context of multiple agents cooperating to solve a problem, there is an optimum number of agents required to maximise speed of learning whilst minimising computational requirements. An investigation into the learning factors of the Phe-Q agent was carried out. The results are presented in this paper.

Section 2 briefly describes ant foraging behaviour and presents some related work inspired by ant foraging mechanisms. Section 3 describes the Phe-Q learning algorithm. In Section 4, optimisation of the Phe-Q parameters is discussed. Experiments and results obtained in optimising the algorithm are presented in Sections 5 and 6 respectively. In Section 7 the results are analysed and finally the paper concludes in Section 8.

2 Ant behaviour and related work

Ants are able to find the shortest path between the nest and a food source by an auto catalytic process [1, 2, 14]. This process comes about because ants deposit pheromones along the trail as they move along in the search for food or resources to construct a nest. The pheromone evaporates with time nevertheless ants follow a pheromone trail and at a branching point prefer to follow the path with higher concentrations of pheromone. On finding the food source, the ants return laden to the nest depositing more pheromone along the way thus reinforcing the pheromone trail. Ants that have followed the shortest path are quicker to return the nest, reinforcing the pheromone trail at a faster rate than those ants that followed an alternative longer route. Further ants arriving at the branching point choose to follow the path with the higher concentrations of pheromone thus reinforcing even further the pheromone and eventually all ants follow the shortest path. The amount of pheromone secreted is a function of an angle between the path and a line joining the food and nest locations [5]. So far two properties of pheromone secretion were mentioned: aggregation and evaporation [18]. The concentration adds when ants deposit pheromone at the same location, and over time the concentration gradually reduces by evaporation. A third property is diffusion [18]. The pheromone at a location diffuses into neighbouring locations.

Ant behaviour has been researched not only for the understanding of the species but also as an inspiration in developing computational problem-solving systems. A methodology inspired by the ant behaviour was developed in [8, 11, 13]. Some of the mechanisms adopted by foraging ants have been applied to classical NP-hard combinatorial optimisation problems with success. In [10] Ant Colony Optimisation is used to solve the travelling salesman problem, a quadratic assignment problem in [13], the job-shop scheduling problem in [7], communication network routing [6] and the Missionaries and Cannibals problem in [17].

In [12] Gambardella suggests a connection between the ant optimisation algorithm and reinforcement learning (RL) and proposes a family of algorithms (Ant-Q) related to Q-learning. The ant optimisation algorithm is a special case of the Ant-Q family. The merging of Ant foraging mechanisms and reinforcement learning is also described in [15]. Three mechanisms found in ant trail formation were used as exploration strategy in a robot navigation task. In this work as with the Ant-Q algorithm, the information provided by the pheromone is used directly for the action selection mechanism.

Another work inspired by ant behaviour is reported in [21]. It is applied to a multi-robotic environment where the robots transport objects between different locations. Rather than physically laying a trail of synthetic pheromones, the robots communicate path information via shared memory.

3 Pheromone-Q learning

The Phe-Q technique combines Q-Learning [22] and synthetic pheromone, by introducing a belief factor into the update equation. The belief factor is a function of the synthetic pheromone concentration on the trail and reflects the extent to which an agent will take into account the information lay down by other agents from the same co-operating group. Reinforcement learning and synthetic pheromone have previously been combined for action selection [15, 21]. The usefulness of the belief factor is that it allows an agent to selectively make use of implicit communication from other agents where the information may not be reliable due to changes in the environment. Incomplete and uncertain information are critical issues in the design of real world systems, Phe-Q addresses this issue by introducing the belief factor into the update equation for Q-Learning.

The main difference between the Q-learning update equation and the pheromone-Q update equation is the introduction of a belief factor that must also be maximised. The belief factor is a function of synthetic pheromone. The synthetic pheromone ($\Phi(s)$) is a scalar value (where s is a state/cell in a grid) that comprises three components: aggregation, evaporation and diffusion. The pheromone $\Phi(s)$ has two possible discrete values, a value for the pheromone deposited when searching for food and when returning to the nest with food. The belief factor (B) dictates the extent to which an agent believes in the pheromone that it detects. An agent, during early training episodes, will believe to a lesser degree in the pheromone map because all agents are biased towards exploration. The belief factor is given by (1)

$$B(s_{t+1}, a) = \frac{\Phi(s_{t+1})}{\sum_{\sigma \in N_a} \Phi(\sigma)} \tag{1}$$

where $\Phi(s)$ is the pheromone concentration at a cell, s, in the environment and N_a is the set of neighbouring states for a chosen action a. The Q-Learning update equation modified with synthetic pheromone is given by (2)

$$\begin{aligned} \hat{Q}_n(s_t, a) \longleftarrow\ & (1 - \alpha_n)\hat{Q}_{n-1}(s_t, a) + \\ & \alpha_n(r_t + \gamma\prime \cdot max_{a'}(\hat{Q}_{n-1}(s_{t+1}, a') + \\ & \xi B(s_{t+1}, a')) \end{aligned} \tag{2}$$

where the parameter, ξ, is a sigmoid function of time ($epochs \geq 0$). The value of ξ increases as the number of agents successfully accomplish the task at hand. It can be shown that the pheromone-Q update equation converges for a non-deterministic MDP. [1]

4 Optimal choice of parameters

The objective is to automate the procedure of fine tuning the many parameters that influence the speed of convergence of Phe-Q learning. In the following discussion, it is assumed that the Q-learning constants α and γ are already optimised. The parameters that influence the speed of learning are the number of agents, pheromone secretion rate, pheromone diffusion rate, pheromone evaporation rate and the pheromone saturation level. Other variable that were found experimentally to influence the convergence were the coefficients of the sigmoid (function of time) that modulates the belief function. The

[1] The proof follows those of Jakkola [20] and Bertsekas [3], assuming that the value function V depends also on the belief factor B: $V_n(s_{t+1}) = max_a(Q_n(s_{t+1}, a) + \xi B(s_{t+1}, a))$.

pheromone distribution in the grid environment is a function of the number of agents in the grid. It is also a function the diffusion of across cells and the evaporation. The pheromone distribution in turn affects the belief function. From equation (1), the belief value depends on the pheromone saturation level. The fine tuning of the parameters for optimum speed of convergence requires a function to minimise. Convergence was also proven empirically by plotting the root mean square of the error between Q values obtained during successive epochs. The function chosen to minimise was the area under the convergence curve. The graph in Figure 1 shows a convergence curve for a Phe-Q learning agent.

First brute force exhaustive search with coarse data was investigated followed by finer data. The purpose of the exhaustive search was to sample the state space and determine if the state space was convex. The values of the parameters were constrained to reduce the search space. With a relatively small space using the selection of the parameters mentioned above an exhaustive search was found to be feasible.

5 Experimental set-up

Evaluation of the modified updating equation for Phe-Q and comparison with the Q-learning update were presented in [16]. The objective in this paper is to present results of investigations into the fine tuning of the parameters upon which speed of convergence depends.

For the experiments reported the agent environment is a $N \times N$, grid where $N = 20$. Obstacles are placed on the grid within cells, preventing agents from occupying the cell. The obstacles were placed such that the goal state was occluded. The agents are placed at a starting cell (the nest) on the grid. The aim is for the agents to locate the 'food' sources occupying a cell in the grid space and return to the starting cell. The agents move from cell to cell in the four cardinal directions, depositing discrete quantities of pheromone in each cell. The two pheromone values (one associated with search for the food source φ_s and the other associated with the return to the nest φ_n) are parameters to fine tune. The pheromone aggregates in a cell up to to a saturation level, and evaporates at a rate (evaporation rate φ_e) until there is none remaining if the cell pheromone is not replenished by an agent visiting the cell. Equally the pheromone diffuses into neighbouring cells at a rate (diffusion rate φ_d), inversely proportional to the manhattan distance.

Each agent has two goal tasks represented each by a Q-table. One task is to reach the 'food' location, and the second task is to return to the nest. Before release into the 'world' agents have no a-priori knowledge of the environment, nor location of nest or resource, etc. More than one agent can occupy a cell. A cell has associated a pheromone strength, $\Phi \in [0, 255]$. Pheromone is de-coupled from the state at the implementation level so that the size of the state space is an $N \times N$. The grids used (N=10 and 20), result in a state space sufficiently small for a lookup table to be used for the Q values.

The agent receives a reward of 1.0 on completing the tasks. Each experiment consists of a number of agents released into the environment and running in parallel until convergence.

6 Results

Convergence was demonstrated empirically by plotting the Root Means Square (RMS) of the error between successive Q-values against epochs (an epoch is a complete cycle of locating food and returning to the nest). The RMS curves for Phe-Q in Figure 1 show convergence. In a 20x20 grid space, the performance degrades with approximately 30 agents. The graph in Figure 2 shows the RMS curves for increasing number of Phe-Q agents maintaining a constant grid size (for clarity only the RMS curves for 5, 40, and 60 agents are shown on the graph). Between 5 and 20 agents, convergence is comparable for the Phe-Q agents. Above that number, the trend is towards slower convergence.

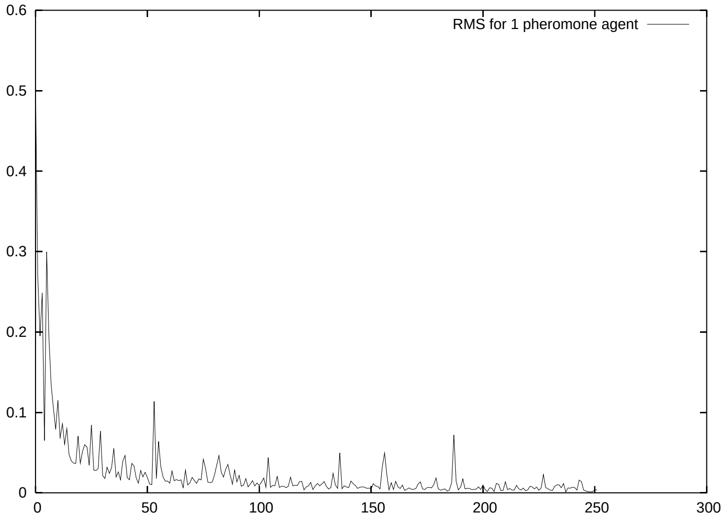

Fig. 1. RMS curve for one Phe-Q agent

As mentioned previously, the objective is to maximise speed of convergence. One way to achieve this is to minimise the area under the RMS curves. The area is a function of independent variables : number of agents, pheromone deposition, diffusion and evaporation. Some constraints are imposed on these variables. These are shown in Table 1. For example, pheromone secretion is a positive scalar, so are evaporation and diffusion. A minimum number of agents in a Multi-Agent system is two. These factors determine the lower bound. To impose upper bounds, results previously obtained were analysed. In a 20 by 20 grid, performance degrades beyond 25 to 30 agents [16]. An upper bound of 20 agents was used. It was also noted that as the number of agents increase, the unit of pheromone secreted (to achieve a performance better than Q-learning) must be reduced. Conversely as the number of agents increase, the unit of evaporation must be increased. Results show that for a given number of agents there is an optimum parameter set. Table 2 below shows the optimum parameter values for 5, 10, 15 and 20 agents.

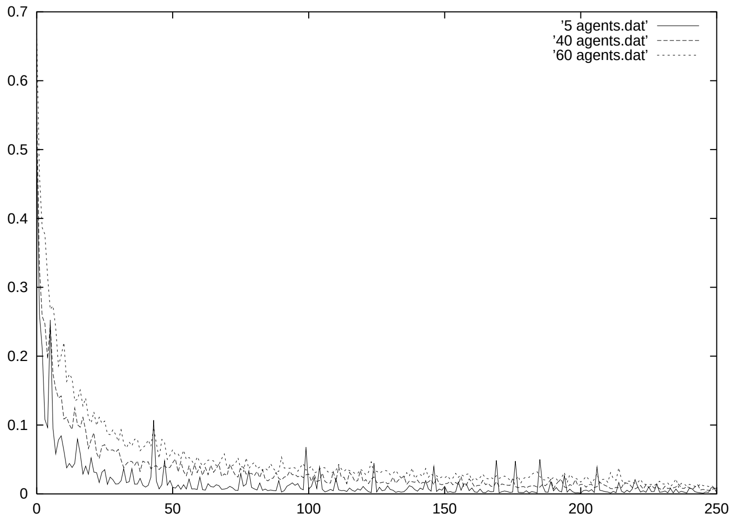

Fig. 2. Performance scaling: 5, 40, 60 agents

Table 1. Constraints on parameters

Parameter	Agent #	φ_s	φ_n	φ_e
Lower bound	2	0	4	0.8
Upper bound	20	1.0	12	2.0

Of the four cases, the system with 5 agents performs better for the particular scenario (grid size, obstacle layout, goal location). For a given problem, too high or too low a pheromone secretion (maintaining a fixed evaporation) produces somewhat degraded results as shown in Figure 3 (slower convergence and noisy Q values). The role of the evaporation is to balance the pheromone secretion. The results in the table show the best performance for each case, the optimal parameters produces (approximately) only slightly worse performance for 10 and 15 agents. The final choice of which parameter set (i.e. number of agents) to use will therefore depend on the computational load. As

Table 2. Optimum values for parameters

Parameter			
Agent #	φ_s	φ_n	φ_e
5	0.8	5.0	0.8
10	0.4	5.0	0.8
15	0.0	5.0	1.6
20	0.0	1.0	0.8

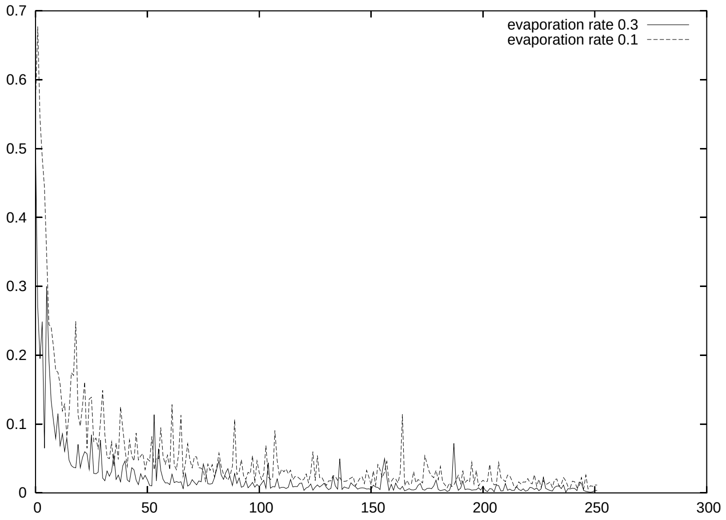

Fig. 3. The role of evaporation

might be expected, the higher the number of agents, the greater the computational load. The next step was to select the optimal parameter set using computational time as a constraint. Table 3 below shows the relative computational times using the parameters shown in table 2.

Table 3. Computational load w.r.t. number of agents

Agent #	Time
5	1.0
10	1.76
15	2.53
20	3.23

7 Discussion

The synthetic pheromone guides the agents. It is implicit communication. The information exchange via pheromone enables the agents to learn quicker however there is a price to pay for this information sharing. Too much information i.e. too high a pheromone deposition rate or too low a pheromone evaporation rate causes not unexpectedly poorer results especially in the earlier learning stages where agents are 'mislead' by other exploring agents. Therefore the parameters choice must be such as to minimise the 'misleading' effect whilst maximising learning. Evaporation is key to reducing agents being mislead in the exploratory phases.

Results show that the required number of agents to optimise learning (achieving a 59% decrease in the area under the RMS curve over Q-learning) is relatively low, this means that the computational load can be maintained low.

8 Conclusions

The work described in this paper set out to investigate the tuning of parameters to achieve optimum learning with the Phe-Q technique in the context of a multi-agent system. By bounding the search space, an exhaustive search was carried. Results confirm the relationships between pheromone deposition and evaporation rates and that evaporation is necessary to prevent 'misleading' information to be received by co-operating agents. An important result concerns the number of agents required to achieve optimum learning. This number is lower than previously expected. An important factor since an increase in agents greatly increases computational times.

References

1. R. Beckers, J. L. Deneubourg, S. Goss, and J. M. Pasteels. Collective decision making through food recruitment. *Ins. Soc.*, 37:258–267, 1990.
2. R. Beckers, J.L. Deneubourg, and S. Goss. Trails and u-turns in the selection of the shortest path by the ant lasius niger. *Journal of Theoretical Biology*, 159:397–4151, 1992.
3. D.P. Bertsekas and J.N. Tsitsiklis. *Neuro-Dynamic Programming*. Athena Scientific, 1996.
4. E. Bonabeau, M. Dorigo, and G. Theraulaz. *Swarm intelligence, From Natural to Artificial Systems*. Oxford University Press, 1999.
5. M. C. Cammaerts-Tricot. Piste et pheromone attraction chez la fourmi myrmica ruba. *Journal of Computational Physiology*, 88:373–382, 1974.
6. G. Di Caro and M. Dorigo. Antnet: a mobile agents approach to adaptive routing.
7. A. Colorni, M. Dorigo, and V. Maniezzo. Ant system for job-shop scheduling. *Belgian Journal of OR, statistics and computer science*, 34:39–53, 1993.
8. A. Colorni, M. Dorigo, and G. Theraulaz. Distributed optimzation by ant colonies. In *Proceedings First European Conf. on Artificial Life*, pages 134–142, 1991.
9. J.L. Deneubourg and S. Goss. Collective patterns and decision making. *Ethol. Ecol. and Evol.*, 1:295–311, 1993.
10. M. Dorigo and L. M. Gambardella. Ant colony system: A cooperative learning approach to the travelling salesman problem. *IEEE Trans. on Evol. Comp.*, 1:53–66, 1997.
11. M. Dorigo, V. Maniezzo, and A. Colorni. The ant system: Optimization by a colony of cooperatin agents. *IEEE Trans. on Systems, Man, and Cybernetics*, 26:1–13, 1996.
12. L. M. Gambardella and M. Dorigo. Ant-q: A reinforcement learning approach to the traveling salesman problem. In *Proc. 12Th ICML*, pages 252–260, 1995.
13. L. M. Gambardella, E. D. Taillard, and M. Dorigo. Ant colonies for the qap. *Journal of Operational Research society*, 1998.
14. S. Goss, S. Aron, J.L. Deneubourg, and J. M. Pasteels. Self-organized shorcuts in the argentine ants. *Naturwissenschaften*, pages 579–581, 1989.
15. L. R. Leerink, S. R. Schultz, and M. A. Jabri. A reinforcement learning exploration strategy based on ant foraging mechanisms. In *Proc. 6Th Australian Conference on Neural Nets*, 1995.

16. N. Monekosso and P. Remagnino. Phe-q: A pheromone based q-learning. In *AI2001:Advances in Artificial Intelligence, 14Th Australian Joint Conf. on A.I.*, pages 345–355, 2001.
17. H. Van Dyke Parunak and S. Brueckner. Ant-like missionnaries and cannibals: Synthetic pheromones for distributed motion control. In *Proc. of ICMAS'00*, 2000.
18. H. Van Dyke Parunak, S. Brueckner, J. Sauter, and J. Posdamer. Mechanisms and military applications for synthetic pheromones. In *Proc. 5Th International Conference Autonomous Agents, Montreal, Canada*, 2001.
19. R. S. Sutton and A.G. Barto. *Reinforcement Learning*. MIT Press, 1998.
20. T.Jaakkola, M.I.Jordan, and S.P.Singh. On the convergence of stochastic iterative dynamic programming algorithms. *Neural Computation*, 6:1185–1201, 1994.
21. R. T. Vaughan, K. Stoy, G. S. Sukhatme, and M. J. Mataric. Whistling in the dark: Cooperative trail following in uncertain localization space. In *Proc. 4Th International Conference on Autonomous Agents, Barcelona, Spain*, 2000.
22. C. J. C. H. Watkins. *Learning with delayed rewards*. PhD thesis, University of Cambridge, 1989.