

Technical report 11-012

# **Convergence Analysis of Ant Colony Learning\***

J. van Ast, R. Babuška, and B. De Schutter

*To cite this work, please refer to the published version:*

J. van Ast, R. Babuška, and B. De Schutter, “Convergence analysis of ant colony learning,” *Proceedings of the 18th IFAC World Congress*, Milan, Italy, pp. 14693–14698, Aug.–Sept. 2011. doi:[10.3182/20110828-6-IT-1002.01533](https://doi.org/10.3182/20110828-6-IT-1002.01533)

Delft Center for Systems and Control  
Delft University of Technology  
Mekelweg 2, 2628 CD Delft  
The Netherlands  
phone: +31-15-278.24.73 (secretary)  
URL: <https://www.dcsc.tudelft.nl>

---

\* This report can also be downloaded via <https://dpub.eu/11-012>

# Convergence Analysis of Ant Colony Learning

Jelmer van Ast\* Robert Babuška\* Bart De Schutter\*

\* Delft Center for Systems and Control of the Delft University of Technology,  
Mekelweg 2, 2628 CD Delft, The Netherlands (e-mail: j.m.vanast@tudelft.nl,  
r.babuska@tudelft.nl, b.deschutter@tudelft.nl)

---

**Abstract:** In this paper, we study the convergence of the pheromone levels of Ant Colony Learning (ACL) in the setting of discrete state spaces and noiseless state transitions. ACL is a multi-agent approach for learning control policies that combines some of the principles found in ant colony optimization and reinforcement learning. Convergence of the pheromone levels in expected value is a necessary requirement for the convergence of the learning process to optimal control policies. In this paper, we derive upper and lower bounds for the pheromone levels and relate those to the learning parameters and the number of ants used in the algorithm. We also derive upper and lower bounds on the expected value of the pheromone levels.

*Keywords:* Evolutionary algorithms in control and identification; Reinforcement learning control

---

## 1. INTRODUCTION

Ant Colony Learning (ACL) is an algorithmic framework for the automatic learning of control policies for non-linear systems with continuous or discrete state spaces (van Ast et al., 2009, 2010). The framework is based on the Ant Colony Optimization (ACO) class of algorithms, which is inspired by the foraging behavior of ants (Dorigo and Blum, 2005). In ACL, a collection of agents – called ants – jointly interact with the system at hand in order to find an optimal control policy, i.e., an optimal mapping between states and actions. Through the stigmergic interaction by means of pheromones, the ants are guided by each other’s experience towards better control policies. In discrete time, the control policy will lead to a sequence of state-action pairs starting in a given initial state and terminating with the action that brings the state of the system to the desired value. A sequence of state-action pairs is called a solution. A predefined cost function evaluates solutions and a solution is then called optimal if it has the lowest cost compared to the cost of all possible solution trajectories from the initial state to the goal state. A policy is called optimal if for all states in the state space the solutions are optimal.

In this paper we present a theoretical study on the convergence of the pheromone levels of ACL. This is a step towards a convergence proof of the algorithm to optimal control policies. We derive upper and lower bounds on the pheromone levels, as well as on the expected value on the pheromone levels. Furthermore, we relate those bounds to the learning parameters and the number of ants used in the algorithm. Convergence analysis of ACL is different from the convergence analysis of the  $\text{ACO}_{\text{gb},\tau_{\text{min}}}$  class of ACO algorithms (Stützle and Dorigo, 2002). The convergence proof of Stützle and Dorigo (2002) mainly relies on the global-best update rule, with which the pheromones are only updated if they belong to the best solution found so far. With ACL, all solutions found in a trial are used to update the pheromone levels. This requires a different type of convergence analysis.

This paper is structured as follows. In Section 2, the ACL framework is presented. Section 3 presents a closed expression

for the total pheromone update that takes place during a trial of the algorithm. In Section 4 upper and lower bounds are derived for the pheromone levels. Based on these results, the upper and lower bounds on the expected value of the pheromone levels is derived in Section 5. The paper concludes with Section 6.

## 2. ANT COLONY LEARNING

### 2.1 The Optimal Policy Learning Problem

Assume that we have a nonlinear dynamic system, characterized by a discrete-valued state vector  $\mathbf{q} = [q_1 \ q_2 \ \dots \ q_n]^T \in \mathcal{Q}$ , where  $\mathcal{Q}$  has a finite number of elements. Also assume that the system can be controlled by an input  $\mathbf{u} \in \mathcal{U}$  that can only take a finite number of values and that the state can be measured at discrete time steps, with a sampling time  $T_s$  with  $t$  the discrete time index. The sampled system is denoted as:

$$\mathbf{q}(t+1) \sim \mathbf{p}(\mathbf{q}(t), \mathbf{u}(t)),$$

with  $\mathbf{p}$  a probability distribution function over the state-action space. The nonlinear mapping function  $\mathbf{h}$  is called the control policy:

$$\mathbf{u}(t) = \mathbf{h}(\mathbf{q}(t)).$$

The optimal control problem we consider is to control the system from any given initial state  $\mathbf{q}(0) = \mathbf{q}_0$  to a desired goal state  $\mathbf{q}(t) = \mathbf{q}_g$  in at most  $t \leq T$  steps and in an optimal way, where optimality is defined by minimizing a certain cost function:

$$J(s) = J(\tilde{\mathbf{q}}, \tilde{\mathbf{u}}), \quad (1)$$

with  $s$  a solution and  $\tilde{\mathbf{q}} = \mathbf{q}(1), \dots, \mathbf{q}(T)$  and  $\tilde{\mathbf{u}} = \mathbf{u}(0), \dots, \mathbf{u}(T-1)$  respectively the sequences of states and actions in that solution. The problem is to find a control policy that, when applied to the system in  $\mathbf{q}_0$ , results in a sequence of state-action pairs  $(\mathbf{u}(0), \mathbf{q}(1)), (\mathbf{u}(1), \mathbf{q}(2)), \dots, (\mathbf{u}(T-1), \mathbf{q}(T))$  that minimizes this cost function. In our case, we aim at finding control policies for non-linear systems, which in general cannot be derived analytically from the system description and the cost function.

The cost function must satisfy for any solution  $s$ :

$$0 < \frac{\tau_0}{\rho} \leq J_{\max}^{-1} \leq J^{-1}(s) \leq J_{\min}^{-1}, \quad (2)$$

with  $J_{\max} = \max_s J(s)$  and  $J_{\min} = \min_s J(s)$  respectively the largest and smallest possible value of the cost function. The parameter  $\tau_0$  represents the initial value of the pheromone levels and  $\rho$  denoted the global pheromone decay rate. These parameters will be further explained in Section 2.2 and Section 2.5 respectively. Note that this requirement is not at all restrictive, since adding a constant  $\frac{\tau_0}{\rho}$  to the cost function renders this requirement satisfied, without changing the optimal solution. Also note that we can trivially extend the optimal control problem that we consider here to include a set of goal states, denoted by  $\mathcal{Q}_g$ . In that case, we can include a single virtual goal state to which all states  $\mathbf{q}_g \in \mathcal{Q}_g$  lead with probability one and a cost of zero.

## 2.2 Outline of the ACL Algorithm

The general outline of the ACL algorithm is as follows. Initially, all  $M$  ants are distributed randomly over the state space of the system and the pheromone levels  $\tau_{\mathbf{qu}}$  associated with each state-action pair  $(\mathbf{q}, \mathbf{u})$  are set to an initial value  $\tau_{\mathbf{qu}}(0) = \tau_0$ . In what is called a *trial*, all ants make interaction steps with the system. First, they decide based on the pheromone levels which action to perform, after which they apply this action to their own copy of the system. They store the state-action pair to their personal record, called their partial solution  $s_p$  and the pheromone level at that state-action pair  $\tau_{\mathbf{qu}}$  is annealed, according to (4), through the local pheromone update. Each copy of the system responds to the input by changing its state, after which the ants repeat the process by choosing a new action until they reach the goal state  $\mathbf{q}_g$ , which terminates the trial. After all ants have terminated their trial, or after a predefined number of trials, all partial solutions are added to the multiset  $\mathcal{S}_{\text{trial}}$ . This set is used in the global pheromone update step, where all solutions in the set are evaluated over the cost function, and the state-action pairs contained in the solutions receive a pheromone update accordingly.

From the output of the algorithm, which are the pheromone values, the control policy can be derived. In the following, we explicitly distinguish between the steps in the inner loop and the steps in the outer loop of the algorithm. In the inner loop, the iterations are indexed by  $t$ , while in the outer loop, the iterations are indexed by  $k$ . In order to understand the timing of the pheromone updates unambiguously, the pheromone variables in the inner loop receive the superscript ‘‘local’’:  $\tau_{\mathbf{qu}}^{\text{local}}$ . Before starting the inner loop, the current pheromone levels are copied to the local pheromone levels:  $\tau_{\mathbf{qu}}^{\text{local}}(0) = \tau_{\mathbf{qu}}(k)$  for all state-action pairs. The first step in the inner loop is the selection of the action.

## 2.3 Action Selection

In the action selection step, each ant  $c$  determines which action to apply to the system in a given state  $\mathbf{q}_c$ . There are various possible forms for the action selection, but the analysis in this paper is based on the  $\epsilon$ -greedy action selection rule. Here, the amount of exploration is kept constant, due to the inclusion of an explicit exploration probability  $\epsilon$ :

$$\mathbf{u}_c = \begin{cases} \arg \max_{\ell \in \mathcal{U}_{\mathbf{q}_c}} \left( \tau_{\mathbf{q}_c \ell}^{\text{local}}(t) \right) & \text{with probability } 1 - \epsilon \\ \text{random}(\mathcal{U}_{\mathbf{q}_c}) & \text{with probability } \epsilon, \end{cases} \quad (3)$$

where  $\text{random}(\mathcal{U}_{\mathbf{q}_c})$  denotes the random selection of an action from the action set  $\mathcal{U}_{\mathbf{q}_c}$  in state  $\mathbf{q}_c$  using a uniform distribution.

## 2.4 Local Pheromone Update

After every step, each ant  $c$  performs a local pheromone update for the  $(\mathbf{q}_c, \mathbf{u}_c)$ -pair just visited:

$$\tau_{\mathbf{q}_c \mathbf{u}_c}^{\text{local}}(t+1) = (1 - \gamma) \tau_{\mathbf{q}_c \mathbf{u}_c}^{\text{local}}(t) + \gamma \tau_0, \quad (4)$$

with  $\gamma \in [0, 1)$  the local pheromone decay rate. The purpose of the local pheromone update is to stimulate exploration of the state-action space, by making it less attractive for an ant to choose the same action in a certain state as its predecessor. When all ants have reached the goal, or when the inner loop has timed-out, the algorithm continues with the global pheromone step in the outer loop.

## 2.5 Global Pheromone Update

After completion of the trial (which, let us assume, happens when  $t = T$ ), the pheromone levels are updated according to the following global pheromone update step:

$$\tau_{\mathbf{qu}}(k+1) = (1 - \rho) \tau_{\mathbf{qu}}^{\text{local}}(T) + \rho \sum_{\substack{s \in \mathcal{S}_{\text{trial}}(k) \\ (\mathbf{q}, \mathbf{u}) \in s}} J^{-1}(s),$$

$$\forall (\mathbf{q}, \mathbf{u}) : \exists s \in \mathcal{S}_{\text{trial}}(k) : (\mathbf{q}, \mathbf{u}) \in s, \quad (5)$$

with  $\mathcal{S}_{\text{trial}}$  the multiset of all candidate solutions found in the trial and  $\rho \in (0, 1]$  the global pheromone decay rate. Note that elitism in the global pheromone update is not possible, since the best solution would then always be the solution starting just prior to the goal state with the action taking the system to the goal state immediately. Since we aim at learning optimal control policies from any initial state, we must also include every solution found in the update. The pheromone deposit is equal to  $J^{-1}(s) = J^{-1}(\tilde{\mathbf{q}}, \tilde{\mathbf{u}})$ , the inverse of the cost function over the sequence of discrete state-action pairs in  $s$ . Note that minimizing the cost corresponds to maximizing the pheromone levels corresponding to the optimal solution. After the global pheromone update, the algorithm continues for  $k+1$  at the start of the outer loop until the maximal number of trials have taken place (i.e., when  $k = K$ ).

## 2.6 Control Policy

The control policy can be extracted from the pheromone levels as follows:

$$\mathbf{u} = \mathbf{h}(\mathbf{q}) = \arg \max_{\ell \in \mathcal{U}_{\mathbf{q}}} (\tau_{\mathbf{q}\ell}), \quad (6)$$

in which ties are broken randomly. This equation means that the control policy assigns the action to a given state that maximizes the associated pheromone levels.

## 3. TOTAL PHEROMONE UPDATE

Let us now derive an expression for the total effect of the pheromone updates during one trial. At the start of a trial  $k$ , the current pheromone levels are copied to the local pheromone levels,  $\tau_{\mathbf{qu}}^{\text{local}}(0) = \tau_{\mathbf{qu}}(k)$  for all  $(\mathbf{q}, \mathbf{u})$ -pairs. In the first

trial, when  $\tau_{\mathbf{q}\mathbf{u}}^{\text{local}}(0) = \tau_{\mathbf{q}\mathbf{u}}(0) = \tau_0$  for all  $(\mathbf{q}, \mathbf{u})$ -pairs, the local pheromone update (4) has no effect. Only after some pheromone variables have received a pheromone deposit from the global pheromone update (5), these pheromone levels can become larger than  $\tau_0$ . Let us assume that the trial is ended when  $t = \bar{T}$  and that  $M_{\mathbf{q}\mathbf{u}}(k)$  ants have visited the  $(\mathbf{q}, \mathbf{u})$ -pair in the given trial. It is then easy to verify that the pheromone levels have then been updated according to:

$$\begin{aligned}\tau_{\mathbf{q}\mathbf{u}}^{\text{local}}(T) &= (1 - \gamma)^{M_{\mathbf{q}\mathbf{u}}(k)}(\tau_{\mathbf{q}\mathbf{u}}^{\text{local}}(0) - \tau_0) + \tau_0 \\ &= (1 - \gamma)^{M_{\mathbf{q}\mathbf{u}}(k)}(\tau_{\mathbf{q}\mathbf{u}}(k) - \tau_0) + \tau_0.\end{aligned}\quad (7)$$

The global pheromone update is applied to  $\tau_{\mathbf{q}\mathbf{u}}^{\text{local}}(T)$  at the end of a trial, if  $(\mathbf{q}, \mathbf{u})$  is an element of the solution of one or more ants. We can aggregate (7) and (5) to get an expression for the total pheromone update at the end of a trial:

$$\begin{aligned}\tau_{\mathbf{q}\mathbf{u}}(k+1) &= (1 - \rho) \left\{ (1 - \gamma)^{M_{\mathbf{q}\mathbf{u}}(k)}(\tau_{\mathbf{q}\mathbf{u}}(k) - \tau_0) + \tau_0 \right\} \\ &\quad + \rho \sum_{\substack{s \in \mathcal{S}_{\text{trial}}(k): \\ (\mathbf{q}, \mathbf{u}) \in s}} J^{-1}(s), \quad \text{if } M_{\mathbf{q}\mathbf{u}}(k) > 0,\end{aligned}\quad (8)$$

$$\tau_{\mathbf{q}\mathbf{u}}(k+1) = \tau_{\mathbf{q}\mathbf{u}}(k), \quad \text{otherwise,} \quad (9)$$

with  $M_{\mathbf{q}\mathbf{u}}(k)$  the number of ants that have visited  $(\mathbf{q}, \mathbf{u})$  during trial  $k$ . From (8) - (9) we can see that:

- (1) If a  $(\mathbf{q}, \mathbf{u})$  is *not* visited by any of the ants in a given trial, the pheromone level  $\tau_{\mathbf{q}\mathbf{u}}$  will not be updated in that trial.
- (2) If a  $(\mathbf{q}, \mathbf{u})$  is visited by one or more ants in a given trial,  $\tau_{\mathbf{q}\mathbf{u}}$  will be updated in that trial, and may increase or decrease in value depending on the pheromone deposits of the ants and the values of  $\gamma$  and  $\rho$ .

By introducing  $\kappa$  as the counter of the number of trials in which the pheromone level of a state-action pair  $(\mathbf{q}, \mathbf{u})$  receives a global pheromone update, we can write (8) - (9) as:

$$\begin{aligned}\tau_{\mathbf{q}\mathbf{u}}^{\text{upd}}(\kappa+1) &= (1 - \rho) \left\{ (1 - \gamma)^{M_{\mathbf{q}\mathbf{u}}(\kappa)}(\tau_{\mathbf{q}\mathbf{u}}^{\text{upd}}(\kappa) - \tau_0) + \tau_0 \right\} \\ &\quad + \rho \sum_{\substack{s \in \mathcal{S}_{\text{trial}}(\kappa): \\ (\mathbf{q}, \mathbf{u}) \in s}} J^{-1}(s),\end{aligned}\quad (10)$$

in which the superscript ‘‘upd’’ is used to avoid confusion between pheromone updates indexed with  $k$  and  $\kappa$ .

#### 4. BOUNDS ON THE PHEROMONE LEVELS

We will proceed our analysis by deriving lower and upper bounds for the pheromone levels. Starting with the lower bound, we prove the following proposition:

*Proposition 4.1.* The lower bound on the pheromone levels is  $\tau_0$ .

**Proof.** By induction, we can show that a pheromone level can never become smaller than  $\tau_0$  when using the pheromone update expression from (10):

$$\begin{aligned}\tau_{\mathbf{q}\mathbf{u}}^{\text{upd}}(0) &= \tau_0 \\ \tau_{\mathbf{q}\mathbf{u}}^{\text{upd}}(\kappa+1) &= (1 - \rho) \left\{ (1 - \gamma)^{M_{\mathbf{q}\mathbf{u}}(\kappa)}(\tau_{\mathbf{q}\mathbf{u}}^{\text{upd}}(\kappa) - \tau_0) + \tau_0 \right\} \\ &\quad + \rho \sum_{\substack{s \in \mathcal{S}_{\text{trial}}(\kappa): \\ (\mathbf{q}, \mathbf{u}) \in s}} J^{-1}(s) \\ &\geq (1 - \rho) \left\{ \underbrace{(1 - \gamma)^{M_{\mathbf{q}\mathbf{u}}(\kappa)}(\tau_{\mathbf{q}\mathbf{u}}^{\text{upd}}(\kappa) - \tau_0) + \tau_0}_{\geq 0 \text{ by induction}} \right\} \\ &\quad + \rho M_{\mathbf{q}\mathbf{u}}(\kappa) J_{\max}^{-1} \\ &\geq (1 - \rho)\tau_0 + \rho J_{\max}^{-1} \geq (1 - \rho)\tau_0 + \tau_0 \geq \tau_0,\end{aligned}$$

in which we have used the condition from (2).

In order to derive the upper bound on the pheromone levels, we use the following lemma:

*Lemma 4.2.* Consider a first-order scalar difference equation:

$$y(k+1) = ay(k) + b,$$

with  $a \in [0, 1)$ ,  $b \in \mathbb{R}$ , and an initial point  $y(0)$ . If  $y(0) \leq \frac{b}{1-a}$ , then  $y(k)$  is non-decreasing with the final value:

$$\lim_{k \rightarrow \infty} y(k) = \frac{b}{1-a}.$$

**Proof.** The final value follows trivially from the final value theorem of the  $z$ -transform (Åström and Wittenmark, 1990). We can prove that the sequence  $y(k)$  is non-decreasing as follows. Using:

$$y(k) = a^k y(0) + (a^{k-1} + \dots + a + 1)b$$

$$y(k+1) = a^{k+1} y(0) + (a^k + a^{k-1} + \dots + a + 1)b$$

we can derive the following expression for the difference between two consecutive time steps of this difference equation:

$$\begin{aligned}y(k+1) - y(k) &= (a^{k+1} - a^k)y(0) + a^k b \\ &= a^k (b + ay(0) - y(0)).\end{aligned}$$

Since  $a \in [0, 1)$ , we have  $a^k \geq 0$ . So, in order to make  $y(k)$  non-decreasing, we need that:  $(b + ay(0) - y(0)) \geq 0$ , which is equal to requiring that  $y(0) \leq \frac{b}{1-a}$ .

Using this lemma, we can prove the following proposition:

*Proposition 4.3.* For any  $(\mathbf{q}, \mathbf{u})$ -pair, the pheromone levels are bounded from above:

$$\tau_{\mathbf{q}\mathbf{u}}^{\text{upd}}(\kappa) \leq \frac{\beta_{\text{upper}} + \rho M J_{\min}^{-1}}{1 - \alpha_{\text{upper}}},$$

with

$$\alpha_{\text{upper}} = (1 - \rho)(1 - \gamma),$$

$$\beta_{\text{upper}} = (1 - \rho) [(1 - \gamma)^M (-\tau_0) + \tau_0],$$

and with  $J_{\min} = \min_s J(s)$  the smallest possible value of the cost function. Note that for this theoretical analysis, it is not necessary to know the value of  $J_{\min}$ .

**Proof.** Considering a given  $(\mathbf{q}, \mathbf{u})$ -pair, let us rewrite (10) as follows:

$$\begin{aligned}\tau_{\mathbf{q}\mathbf{u}}^{\text{upd}}(\kappa+1) &= (1 - \rho)(1 - \gamma)^{M_{\mathbf{q}\mathbf{u}}(\kappa)} \tau_{\mathbf{q}\mathbf{u}}^{\text{upd}}(\kappa) \\ &\quad + (1 - \rho) \left[ (1 - \gamma)^{M_{\mathbf{q}\mathbf{u}}(\kappa)} (-\tau_0) + \tau_0 \right] \\ &\quad + \rho \sum_{\substack{s \in \mathcal{S}_{\text{trial}}(\kappa): \\ (\mathbf{q}, \mathbf{u}) \in s}} J^{-1}(s).\end{aligned}$$

This equation is of the form:

$$\tau_{\mathbf{qu}}^{\text{upd}}(\kappa + 1) = \alpha(\kappa)\tau_{\mathbf{qu}}^{\text{upd}}(\kappa) + \beta(\kappa) + \delta(\kappa).$$

Let us now introduce  $\theta_{\mathbf{qu}}(\kappa)$ , which satisfies the following difference equation:

$$\theta_{\mathbf{qu}}(\kappa + 1) = \alpha_{\text{upper}}\theta_{\mathbf{qu}}(\kappa) + \beta_{\text{upper}} + \delta_{\text{upper}},$$

in which  $\alpha_{\text{upper}}$ ,  $\beta_{\text{upper}}$ , and  $\delta_{\text{upper}}$  are upper bounds of  $\alpha(\kappa)$ ,  $\beta(\kappa)$ , and  $\delta(\kappa)$  respectively:

$$\begin{aligned} \alpha(\kappa) &\leq \alpha_{\text{upper}} = (1 - \rho)(1 - \gamma), \\ \beta(\kappa) &\leq \beta_{\text{upper}} = (1 - \rho) [(1 - \gamma)^M(-\tau_0) + \tau_0], \\ \delta(\kappa) &\leq \delta_{\text{upper}} = \rho M J_{\min}^{-1}. \end{aligned}$$

The upper bound for  $\alpha(\kappa)$  is obtained by taking  $M_{\mathbf{qu}}(\kappa) = 1$  for all  $\kappa$ , while the upper bounds for  $\beta(\kappa)$  and  $\delta(\kappa)$  are obtained for  $M_{\mathbf{qu}}(\kappa) = M$  for all  $\kappa$ . We take the same initial values for  $\tau_{\mathbf{qu}}^{\text{upd}}$  and  $\theta_{\mathbf{qu}}$ , so  $\theta_{\mathbf{qu}}(0) = \tau_0$ . Using Lemma 4.2, we can now show that  $\theta_{\mathbf{qu}}(\kappa)$  is a non-decreasing function of  $\kappa$ . We immediately see that  $\alpha_{\text{upper}} \geq 0$  and we must show that:

$$\frac{\beta_{\text{upper}} + \delta_{\text{upper}}}{1 - \alpha_{\text{upper}}} \geq \theta_{\mathbf{qu}}(0) = \tau_0.$$

This condition is satisfied if

$$\frac{(1 - \rho) [(1 - \gamma)^M(-\tau_0) + \tau_0] + \rho M J_{\min}^{-1}}{1 - (1 - \rho)(1 - \gamma)} \geq \tau_0,$$

or

$$(1 - \rho) [(1 - \gamma)^M(-\tau_0) + \tau_0] + \rho M J_{\min}^{-1} - \tau_0 + (1 - \rho)(1 - \gamma)\tau_0 \geq 0.$$

Recalling the ranges for  $\rho$ , viz.  $(0, 1]$  and for  $\gamma$ , viz.  $[0, 1)$ , we can show that the latter inequality holds, as follows:

$$\begin{aligned} &\underbrace{(1 - \rho) \left[ \underbrace{(1 - \gamma)^M(-\tau_0) + \tau_0}_{\geq 0} \right]}_{\geq 0} + \underbrace{\rho M J_{\min}^{-1} - \tau_0}_{\geq \tau_0} \\ &+ \underbrace{(1 - \rho)(1 - \gamma)\tau_0}_{\geq 0} \geq 0. \end{aligned}$$

Moreover, by induction it follows that  $\tau_{\mathbf{qu}}^{\text{upd}}(\kappa) \leq \theta_{\mathbf{qu}}(\kappa)$ :

$$\begin{aligned} \tau_{\mathbf{qu}}^{\text{upd}}(0) &= \theta_{\mathbf{qu}}(0) \\ \tau_{\mathbf{qu}}^{\text{upd}}(\kappa + 1) &= \alpha(\kappa)\tau_{\mathbf{qu}}^{\text{upd}}(\kappa) + \beta(\kappa) + \delta(\kappa) \\ &\leq \alpha_{\text{upper}}\theta_{\mathbf{qu}}(\kappa) + \beta_{\text{upper}} + \delta_{\text{upper}} \\ &= \theta_{\mathbf{qu}}(\kappa + 1). \end{aligned}$$

We can use Lemma 4.2 to get  $\lim_{\kappa \rightarrow \infty} \theta_{\mathbf{qu}}(\kappa) = \frac{\beta_{\text{upper}} + \delta_{\text{upper}}}{1 - \alpha_{\text{upper}}}$ .

Since we have shown that  $\theta_{\mathbf{qu}}(\kappa)$  is non-decreasing, we have now arrived at the conclusion that:

$$\tau_{\mathbf{qu}}^{\text{upd}}(\kappa) \leq \theta_{\mathbf{qu}}(\kappa) \leq \frac{\beta_{\text{upper}} + \rho M J_{\min}^{-1}}{1 - \alpha_{\text{upper}}}.$$

Note that this upper bound is only tight for  $M = 1$ .

## 5. BOUNDS ON THE EXPECTED VALUE OF THE PHEROMONE LEVELS

The bounds derived in the previous section are useful as they give us information about the range in which the pheromone levels reside. Moreover, the bounds give the relations between the global and local pheromone decay rates  $\rho$  and  $\gamma$ , the number of ants  $M$ , and the initial value of the pheromone levels  $\tau_0$ .

However, two shortcomings of these bounds prevent us from drawing conclusions about convergence of the algorithm:

- (1) Our upper bound is not tight for  $M > 1$ .
- (2) Our upper bound is the same for all pheromone levels, associated with all  $(\mathbf{q}, \mathbf{u})$ -pairs.

Especially the second issue prevents us from analyzing whether and when the pheromone levels associated with the optimal state-action pairs become larger than the pheromone levels associated with suboptimal state-action pairs. In this section, our aim is to find expressions for the evolution of individual pheromone levels. However, two factors in the algorithm in particular complicate such an analysis:

- (1) The total pheromone update from (10) contains  $M_{\mathbf{qu}}(\kappa)$ , the number of ants that have visited the  $(\mathbf{q}, \mathbf{u})$ -pair in trial  $\kappa$ , which is dependent on many unknown factors, such as the other pheromone levels and the exploration process.
- (2) The total pheromone update of  $\tau_{\mathbf{qu}}^{\text{upd}}(\kappa)$  depends on the pheromone deposits  $J^{-1}(s)$  from all solutions found in trial  $\kappa$ . The update of a particular pheromone level thus depends on all state-action pairs prior to  $(\mathbf{q}, \mathbf{u})$  and all state-action pairs following  $(\mathbf{q}, \mathbf{u})$ , which also depends on many unknown factors, such as the other pheromone levels and the exploration process.

There are too many uncertainties involved in the algorithm in order to find a closed expression of the final pheromone levels. This is inherent to learning algorithms that contain random variables, such as exploration, and in which the *credit assignment*, such as the distribution of rewards in reinforcement learning (Sutton and Barto, 1998), or the distribution of the pheromone deposits in ACL, depends on a series of decision variables.

In the following, we eliminate the uncertainty arising from exploration by looking at the expected value of the pheromone levels. We eliminate the uncertainty arising from the state-action pairs prior to  $(\mathbf{q}, \mathbf{u})$  and the state-action pairs following  $(\mathbf{q}, \mathbf{u})$  by considering the situation depicted in Figure 1.

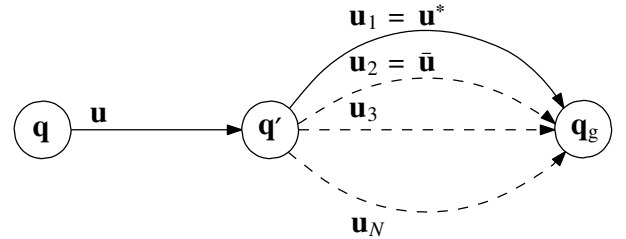


Fig. 1. From a state  $\mathbf{q}$ , the action  $\mathbf{u}$  will take the system to another state  $\mathbf{q}'$ , from which there are  $N$  possible actions. The action  $\mathbf{u}^*$  in this state will bring the system to the goal state optimally. The action  $\bar{\mathbf{u}}$  does so sub-optimally, but still with a lower cost than the other actions.

Here, we regard the  $M_{\mathbf{qu}}(\kappa)$  ants to start in state  $\mathbf{q}$  and all choose the action  $\mathbf{u}$ . All ants are taken to  $\mathbf{q}'$  after which they can choose between  $N$  possible actions. Without loss of generality, we assume that the action  $\mathbf{u}_1 = \mathbf{u}^*$  then takes an ant to the goal state immediately and with the lowest cost compared to the other available actions. It is thus considered to be the optimal action. The action  $\mathbf{u}_2 = \bar{\mathbf{u}}$  is the second-best action. It takes an ant to the goal state with a higher cost compared to  $\mathbf{u}_1$ , but with a lower cost compared to all the other actions. Let us also

assume that the inverse of the cost resulting from all possible actions is ordered as follows:

$$\begin{aligned} J_{\min, \mathbf{q}'}^{-1} &= J^{-1}(\mathbf{q}', \mathbf{u}_1 = \mathbf{u}^*) > J^{-1}(\mathbf{q}', \mathbf{u}_2 = \bar{\mathbf{u}}) \\ &= J_{\text{second}, \mathbf{q}'}^{-1} > \dots > J^{-1}(\mathbf{q}', \mathbf{u}_N) = J_{\max, \mathbf{q}'}^{-1} \end{aligned}$$

Here  $J_{\mathbf{q}'}^{-1}$  is a short-hand expression for the cost that results from an action chosen in  $\mathbf{q}'$  and possible other state-action pairs following  $\mathbf{q}'$ . The subscripts “min”, “second”, etc. then denote respectively the lowest, or next to lowest possible cost to resulting in this manner. By definition,  $J^{-1}(\mathbf{q}', \mathbf{u}_1 = \mathbf{u}^*) = J_{\min, \mathbf{q}'}^{-1}$  and  $J^{-1}(\mathbf{q}', \mathbf{u}_N) = J_{\max, \mathbf{q}'}^{-1}$ .

We will analyze the behavior of the expected value of  $\tau_{\mathbf{qu}}^{\text{upd}}(\kappa)$ . Since the cost resulting from the state-action pair  $(\mathbf{q}, \mathbf{u})$  is independent from the action chosen in  $\mathbf{q}'$ , any constant cost for  $J^{-1}(\mathbf{q}, \mathbf{u})$  will do for our analysis. Without loss of generality and for the sake of simplicity, we take  $J^{-1}(\mathbf{q}, \mathbf{u}) = 0$ , although formally this is impossible, since  $J^{-1}(\mathbf{q}, \mathbf{u}) \geq J_{\max}^{-1} > 0$  for any  $(\mathbf{q}, \mathbf{u})$ -pair. Note that it thus also holds that  $J_{\min, \mathbf{q}}^{-1} = J_{\min, \mathbf{q}'}^{-1}$ ,  $J_{\text{second}, \mathbf{q}}^{-1} = J_{\text{second}, \mathbf{q}'}^{-1}$ , etc. We assume in this section that the optimal action  $\mathbf{u}^*$  from  $\mathbf{q}'$  is currently also associated with the highest pheromone level and is thus also designated to be optimal. During learning, this does not have to be the case, since other actions may be associated with higher pheromone levels and  $\mathbf{u}^*$  is thus not yet known to be the optimal action.

Computing the expected value of  $\tau_{\mathbf{qu}}^{\text{upd}}(\kappa)$  involves taking the expected value of the number of ants  $M_{\mathbf{qu}}(\kappa)$  in the exponent, which severely complicates deriving an analytical expression for the expected value of  $\tau_{\mathbf{qu}}^{\text{upd}}(\kappa)$ . We must thus shift our aim by choosing to derive upper and lower bounds on the expected value of  $\tau_{\mathbf{qu}}^{\text{upd}}(\kappa)$  instead.

*Proposition 5.1.* For any state-action pair  $(\mathbf{q}, \mathbf{u})$ , the expected value of the pheromone levels is bounded from above:

$$E[\tau_{\mathbf{qu}}^{\text{upd}}(\kappa)] \leq \frac{\beta_{\text{upper}} + \rho M J_{\text{exp}, \mathbf{q}}^{-1}}{1 - \alpha_{\text{upper}}}, \quad (11)$$

with

$$\begin{aligned} \alpha_{\text{upper}} &= (1 - \rho)(1 - \gamma), \\ \beta_{\text{upper}} &= (1 - \rho) [(1 - \gamma)^M (-\tau_0) + \tau_0], \\ J_{\text{exp}, \mathbf{q}}^{-1} &= (1 - \epsilon) J_{\min, \mathbf{q}}^{-1} + \epsilon J_{\text{avg}, \mathbf{q}}^{-1}, \end{aligned}$$

and  $J_{\text{avg}, \mathbf{q}}^{-1}$  the inverse of the average cost expected to result when moving from  $\mathbf{q}$  to the goal.

**Proof.** When choosing  $\mathbf{u}^*$ , the pheromone level  $\tau_{\mathbf{qu}}^{\text{upd}}$  is increased the most when:

$$\begin{aligned} \tau_{\mathbf{qu}}^{\text{upd}}(\kappa + 1) &= \underbrace{(1 - \rho)(1 - \gamma)}_{\alpha_{\text{upper}}} \tau_{\mathbf{qu}}^{\text{upd}}(\kappa) + \rho M J_{\min, \mathbf{q}}^{-1} \\ &\quad + \underbrace{(1 - \rho) [(1 - \gamma)^M (-\tau_0) + \tau_0]}_{\beta_{\text{upper}}}. \end{aligned}$$

When choosing  $\bar{\mathbf{u}}$ , the pheromone level  $\tau_{\mathbf{qu}}^{\text{upd}}$  is increased the most when:

$$\begin{aligned} \tau_{\mathbf{qu}}^{\text{upd}}(\kappa + 1) &= \underbrace{(1 - \rho)(1 - \gamma)}_{\alpha_{\text{upper}}} \tau_{\mathbf{qu}}^{\text{upd}}(\kappa) + \rho M J_{\text{second}, \mathbf{q}}^{-1} \\ &\quad + \underbrace{(1 - \rho) [(1 - \gamma)^M (-\tau_0) + \tau_0]}_{\beta_{\text{upper}}}. \end{aligned}$$

The largest increase of  $\tau_{\mathbf{qu}}^{\text{upd}}$  for the other actions follows in a similar manner. The probability that the optimal action  $\mathbf{u}^*$  is chosen is:

$$p(\mathbf{u}^*) = 1 - \epsilon + \frac{\epsilon}{N} = 1 - \left( \frac{N-1}{N} \right) \epsilon,$$

namely the probability of not exploring plus the probability of selecting that action while exploring (which is uniformly distributed). The probability of choosing any of the other actions is  $p(\mathbf{u}_i) = \frac{\epsilon}{N}$ , for  $\mathbf{u}_i \neq \mathbf{u}^*$ . The expected value of  $\tau_{\mathbf{qu}}^{\text{upd}}$  when increasing the most can now be computed as follows:

$$\begin{aligned} &E_{\text{upper}}[\tau_{\mathbf{qu}}^{\text{upd}}(\kappa + 1)] \\ &= \left[ 1 - \left( \frac{N-1}{N} \right) \epsilon \right] (\alpha_{\text{upper}} \tau_{\mathbf{qu}}^{\text{upd}}(\kappa) + \beta_{\text{upper}} + \rho M J_{\min, \mathbf{q}}^{-1}) \\ &\quad + \left[ \frac{\epsilon}{N} \right] (\alpha_{\text{upper}} \tau_{\mathbf{qu}}^{\text{upd}}(\kappa) + \beta_{\text{upper}} + \rho M J_{\text{second}, \mathbf{q}}^{-1}) \\ &\quad \vdots \\ &\quad + \left[ \frac{\epsilon}{N} \right] (\alpha_{\text{upper}} \tau_{\mathbf{qu}}^{\text{upd}}(\kappa) + \beta_{\text{upper}} + \rho M J_{\max, \mathbf{q}}^{-1}) \\ &= \alpha_{\text{upper}} \tau_{\mathbf{qu}}^{\text{upd}}(\kappa) + \beta_{\text{upper}} + \left[ 1 - \left( \frac{N-1}{N} \right) \epsilon \right] \rho M J_{\min, \mathbf{q}}^{-1} \\ &\quad + \underbrace{\left[ \frac{\epsilon}{N} \right] \rho M J_{\text{second}, \mathbf{q}}^{-1} + \dots + \left[ \frac{\epsilon}{N} \right] \rho M J_{\max, \mathbf{q}}^{-1}}_{N-1 \text{ terms}} \\ &= \alpha_{\text{upper}} \tau_{\mathbf{qu}}^{\text{upd}}(\kappa) + \beta_{\text{upper}} + [1 - \epsilon] \rho M J_{\min, \mathbf{q}}^{-1} \\ &\quad + \left[ \frac{\epsilon}{N} \right] \rho M (J_{\min, \mathbf{q}}^{-1} + J_{\text{second}, \mathbf{q}}^{-1} + \dots + J_{\max, \mathbf{q}}^{-1}) \\ &= \alpha_{\text{upper}} \tau_{\mathbf{qu}}^{\text{upd}}(\kappa) + \beta_{\text{upper}} + \rho M J_{\text{exp}, \mathbf{q}}^{-1}, \end{aligned}$$

where  $J_{\text{exp}, \mathbf{q}}^{-1} = (1 - \epsilon) J_{\min, \mathbf{q}}^{-1} + \epsilon J_{\text{avg}, \mathbf{q}}^{-1}$  is the expected pheromone deposit on  $\tau_{\mathbf{qu}}^{\text{upd}}$ . Since

$$E_{\text{upper}}[\tau_{\mathbf{qu}}^{\text{upd}}(\kappa)] \geq E[\tau_{\mathbf{qu}}^{\text{upd}}(\kappa)] \quad \text{for all } \kappa,$$

the following difference equation describes the evolution of the upper bound of the expected value of  $\tau_{\mathbf{qu}}^{\text{upd}}(\kappa)$ :

$$\begin{aligned} E_{\text{upper}}[\tau_{\mathbf{qu}}^{\text{upd}}(\kappa + 1)] &= \alpha_{\text{upper}} E_{\text{upper}}[\tau_{\mathbf{qu}}^{\text{upd}}(\kappa)] + \beta_{\text{upper}} \\ &\quad + \rho M J_{\text{exp}, \mathbf{q}}^{-1}. \end{aligned}$$

Since we can show that

$$E_{\text{upper}}[\tau_{\mathbf{qu}}^{\text{upd}}(0)] = \tau_0 \leq \frac{\beta_{\text{upper}} + \rho M J_{\text{exp}, \mathbf{q}}^{-1}}{1 - \alpha_{\text{upper}}},$$

we know from Lemma 4.2 that  $E_{\text{upper}}[\tau_{\mathbf{qu}}^{\text{upd}}(\kappa)]$  is non-decreasing and that:

$$E[\tau_{\mathbf{qu}}^{\text{upd}}(\kappa)] \leq E_{\text{upper}}[\tau_{\mathbf{qu}}^{\text{upd}}(\kappa)] \leq \frac{\beta_{\text{upper}} + \rho M J_{\text{exp}, \mathbf{q}}^{-1}}{1 - \alpha_{\text{upper}}}.$$

Note that  $J_{\text{avg}, \mathbf{q}}^{-1}$  is generally not known, although it might be possible to estimate it. Similar to the upper bound, we can derive a lower bound on the expected value of pheromone levels.

*Proposition 5.2.* For any state-action pair  $(\mathbf{q}, \mathbf{u})$ , in the limit for  $\kappa \rightarrow \infty$ , the expected value of the pheromone levels is bounded from below:

$$\lim_{\kappa \rightarrow \infty} E[\tau_{\mathbf{qu}}^{\text{upd}}(\kappa)] \geq \frac{\beta_{\text{lower}} + \rho J_{\text{exp}, \mathbf{q}}^{-1}}{1 - \alpha_{\text{lower}}}, \quad (12)$$

with

$$\begin{aligned} \alpha_{\text{lower}} &= (1 - \rho)(1 - \gamma)^M, \\ \beta_{\text{lower}} &= (1 - \rho) [(1 - \gamma)^M (-\tau_0) + \tau_0], \\ J_{\text{exp}, \mathbf{q}}^{-1} &= (1 - \epsilon) J_{\min, \mathbf{q}}^{-1} + \epsilon J_{\text{avg}, \mathbf{q}}^{-1}. \end{aligned}$$

**Proof.** When choosing  $\mathbf{u}^*$ ,  $\tau_{\mathbf{qu}}^{\text{upd}}$  is increased the least when:

$$\tau_{\mathbf{qu}}^{\text{upd}}(\kappa + 1) = \underbrace{(1 - \rho)(1 - \gamma)^M}_{\alpha_{\text{lower}}} \tau_{\mathbf{qu}}^{\text{upd}}(\kappa) + \underbrace{(1 - \rho)[(1 - \gamma)(-\tau_0) + \tau_0]}_{\beta_{\text{lower}}} + \rho J_{\text{min}, \mathbf{q}}^{-1},$$

since for all  $\kappa$   $\alpha(\kappa) = (1 - \rho)(1 - \gamma)^{M_{\mathbf{qu}}(\kappa)}$  is the smallest for  $M_{\mathbf{qu}}(\kappa) = M$  and  $\beta(\kappa) = (1 - \rho)[(1 - \gamma)^{M_{\mathbf{qu}}(\kappa)}(-\tau_0) + \tau_0]$  is the smallest for  $M_{\mathbf{qu}}(\kappa) = 1$ . The rest of the proof follows easily along the same lines as the proof of Proposition 5.1.

The expected value of the pheromone levels lies between these bounds,  $E_{\text{lower}}[\tau_{\mathbf{qu}}^{\text{upd}}(\kappa)] \leq E[\tau_{\mathbf{qu}}^{\text{upd}}(\kappa)] \leq E_{\text{upper}}[\tau_{\mathbf{qu}}^{\text{upd}}(\kappa)]$ , for all  $\kappa$ . In the limit for  $\kappa \rightarrow \infty$ , the expected value of the pheromone levels lies between the derived upper and lower bounds:

$$\frac{\beta_{\text{lower}} + \rho J_{\text{exp}, \mathbf{q}}^{-1}}{1 - \alpha_{\text{lower}}} \leq \lim_{\kappa \rightarrow \infty} E[\tau_{\mathbf{qu}}^{\text{upd}}(\kappa)] \leq \frac{\beta_{\text{upper}} + \rho M J_{\text{exp}, \mathbf{q}}^{-1}}{1 - \alpha_{\text{upper}}}.$$

The expressions for  $\alpha(\kappa)$ ,  $\alpha_{\text{upper}}$ ,  $\alpha_{\text{lower}}$ ,  $\beta(\kappa)$ ,  $\beta_{\text{upper}}$ , and  $\beta_{\text{lower}}$  are presented in Table 1.

Table 1. The expressions for  $\alpha(\kappa)$  and  $\beta(\kappa)$  and their bounds.

$\alpha(\kappa)$ and its upper and lower bound
$\alpha(\kappa) = (1 - \rho)(1 - \gamma)^{M_{\mathbf{qu}}(\kappa)}$
$\alpha_{\text{upper}} = (1 - \rho)(1 - \gamma)$
$\alpha_{\text{lower}} = (1 - \rho)(1 - \gamma)^M$
$\beta(\kappa)$ and its upper and lower bound
$\beta(\kappa) = (1 - \rho)[(1 - \gamma)^{M_{\mathbf{qu}}(\kappa)}(-\tau_0) + \tau_0]$
$\beta_{\text{upper}} = (1 - \rho)[(1 - \gamma)^M(-\tau_0) + \tau_0]$
$\beta_{\text{lower}} = (1 - \rho)[(1 - \gamma)(-\tau_0) + \tau_0] = \tau_0 \gamma (1 - \rho)$

## 6. CONCLUSIONS

In this paper, we have analyzed the evolution of the pheromone levels in Ant Colony Learning. In particular, we have studied the situation in which from a state  $\mathbf{q}$ , an action  $\mathbf{u}$  will take the system to another state  $\mathbf{q}'$ , from which there are  $N$  possible actions. There, the pheromone level  $\tau_{\mathbf{qu}}$  was dependent on the action chosen in the state  $\mathbf{q}'$  further “downstream” towards the goal. The goal would be reached after choosing the optimal action  $\mathbf{u}^*$  in  $\mathbf{q}'$ . Because of the unpredictability of the number of ants that visit a particular state, we have derived upper and lower bounds on the pheromone levels, followed by upper and lower bounds on the expected value of the pheromone levels. In future research, we will extend the convergence analysis towards a convergence proof of the algorithm to optimal control policies. The convergence analysis of the pheromone levels, as presented in this paper, is an essential step for this. Future research must also focus on the applicability of the convergence results to real-world control problems, where the structure of the control problem must be taken into account.

## ACKNOWLEDGMENT

This research is financially supported by Senter, Ministry of Economic Affairs of the Netherlands within the BSIK-

ICIS project “Self-Organizing Moving Agents” (grant no. BSIK03024)

## REFERENCES

- Åström, K.J. and Wittenmark, B. (1990). *Computer Controlled Systems – Theory and Design*. Prentice-Hall, Englewood Cliffs, NJ.
- Dorigo, M. and Blum, C. (2005). Ant colony optimization theory: a survey. *Theoretical Computer Science*, 344(2-3), 243–278.
- Stützle, T. and Dorigo, M. (2002). A short convergence proof for a class of ant colony optimization algorithms. *IEEE Transactions on Evolutionary Computation*, 6(4), 358–365.
- Sutton, R.S. and Barto, A.G. (1998). *Reinforcement Learning: An Introduction*. MIT Press, Cambridge, MA.
- van Ast, J.M., Babuška, R., and De Schutter, B. (2009). Novel ant colony optimization approach to optimal control. *International Journal of Intelligent Computing and Cybernetics*, 2(3), 414–434.
- van Ast, J.M., Babuška, R., and De Schutter, B. (2010). Ant colony learning algorithm for optimal control. In R. Babuška and F.C.A. Groen (eds.), *Interactive Collaborative Information Systems*, volume 281 of *Studies in Computational Intelligence*, 155–182. Springer Berlin / Heidelberg.