Minimising Regret in Route Choice (Doctoral Consortium)

Gabriel de O. Ramos
Instituto de Informática
Universidade Federal do Rio Grande do Sul
Porto Alegre, RS, Brazil
goramos@inf.ufrgs.br

ABSTRACT

The use of reinforcement learning (RL) in multiagent scenarios is challenging. I consider the route choice problem, where drivers must choose routes that minimise their travel times. Here, selfish RL-agents must adapt to each others' decisions. In this work, I show how the agents can learn (with performance guarantees) by minimising the regret associated with their decisions, thus achieving the User Equilibrium (UE). Considering the UE is inefficient from a global perspective, I also focus on bridging the gap between the UE and the system optimum. In contrast to previous approaches, this work drops any full knowledge assumption.

Keywords

regret; route choice; multiagent reinforcement learning

1. INTRODUCTION

Reinforcement learning (RL) in multiagent domains is a challenging task. An RL-agent must learn by trial-and-error how to behave within the environment in order to maximise its utility. When multiple agents share a common environment, they must adapt their behaviour to those of others. The problem becomes even harder when agents are selfish and compete for a common resource.

I consider the *route choice problem*, where rational drivers must choose the routes that minimise their travel times. Learning is fundamental here because agents must adapt their choices to account for changing traffic conditions.

An interesting class of multiagent RL techniques comprises the regret minimisation approaches. Roughly, the so-called *external regret* measures how much worse an agent performs on average as compared to his best fixed action in hindsight. Thus, regret minimisation can be seen as an inherent definition on how rational agents behave over time.

The use of regret in route choice and related problems has been widely explored in the literature. Some progress has been made in the online optimisation of routing games [2]. However, as opposed to these approaches, traffic is intrinsically distributed. Multiagent RL fits better here. Nonetheless, within RL, regret has been mainly employed as a performance measure. Unlike previous approaches, I use regret to *guide* the learning process. Some works indeed employ regret as reinforcement signal [4], but assuming that agents have full knowledge of the environment. However, given the selfish nature of traffic, investigating methods for minimising regret in the absence of global information is more challenging. Hence, this research also provides methods for *estimating* regret, dropping any full knowledge assumption.

Another important aspect of route choice refers to the inefficiency of the UE. Ideally, the system optimum (SO) would be preferred, which represents the system at its best operation. In this sense, several works have tried to move the equilibrium towards the SO. Promising results were obtained with difference rewards [1], where the contribution of a single agent to the global outcome is used as reinforcement signal. However, it assumes agents have global knowledge.

Appears in: *Proc. of the 16th International Conference on Autonomous Agents and Multiagent Systems (AAMAS 2017), S. Das, E. Durfee, K. Larson, M. Winikoff (eds.), May 8–12, 2017, São Paulo, Brazil.*

2. THESIS PROPOSAL

My work is motivated by the following questions: (i) can an agent estimate its regret and learn from it? (ii) is it possible to bound agents' performance? (iii) is there a system-efficient equilibrium? To answer these questions, my thesis is divided into two fronts:

Learning from regret. Here, I study how regret can be locally estimated by the agents and how such a value can be used in the learning process. Furthermore, I investigate theoretical performance guarantees. These results are reported on my AAMAS 2017 paper [3].

System-efficient equilibria. The goal here is to bridge the gap between the UE and the SO by incorporating some sort of global performance on the agents' regret. My previous theoretical results shall be extended to these settings. This second front is under development.

3. PROBLEM MODELLING

A road network can be represented by a directed graph, where nodes represent intersections and links represent the roads between intersections. A link's cost is a function of the flow of vehicles on it. In this work, every driver $i \in D$ is a Q-learning agent with an origin, a destination, and a set of K routes $A_i = \{a_1, \ldots, a_K\}$ connecting them. The problem is modelled as a stateless MDP.

4. LEARNING FROM REGRET

In the first topic of my thesis, I investigate how driver-agents can learn using their regret as reinforcement signal. However, note that an agent cannot compute its regret due

to the lack of information regarding its routes' rewards (i.e., it only observes the reward of the taken route). Moreover, regret does not specify how much a particular action contributes to an agent's reward, thus being useless as reinforcement signal. To overcome these limitations, we define the *action regret* and develop a method for the agents to *estimate* it. This work was accepted at AAMAS 2017 [3].

Let $\tilde{r}(a_i^t)$ represent the newest reward experience of agent i for taking action a on time t (either the current reward or the last actually observed one). Agents keep this information for each of their K actions. We can then formulate the *estimated action regret* (as in Equation (1)), which accounts for the average amount lost by agent i up to time T for taking action a (latter term) as compared to the best action regarding its experience (former term). Additionally, the *estimated external regret* can be obtained by considering the observed rewards in the latter term of Equation (1). The main advantage of our formulation is that it can be computed locally by the agents, eliminating the need for a central authority. Consequently, regret can be used to guide the learning process.

$$\tilde{\mathcal{R}}_{i,a}^T = \max_{b_i^t \in A_i} \frac{1}{T} \sum_{t=1}^{T} \tilde{r}(b_i^t) - \frac{1}{T} \sum_{t=1}^{T} \tilde{r}(a_i^t) \quad (1)$$

Using the above formulation, we can define the RL process. At each episode t, each agent $i \in D$ chooses an action $\dot{a}_i^t \in A_i$ using the ϵ-greedy exploration strategy. After taking the chosen action, the agent receives a reward of $r(\dot{a}_i^t)$, which is used to estimate the regret of action $\dot{a}_i^t$ using Equation (1). Finally, the agent updates $Q(\dot{a}_i^t)$ *using the estimated action regret for that action*, as in Equation (2). After each episode, the learning and exploration rates, α and ϵ, are multiplied by decay rates $\lambda \in (0, 1]$ and $\mu \in (0, 1]$, respectively.

$$Q(\dot{a}_i^t) = (1 - \alpha) Q(\dot{a}_i^t) + \alpha \tilde{\mathcal{R}}_{i,\dot{a}_i^t}^t \quad (2)$$

Theoretical results. Due to the limited space, here I concentrate on the big picture of our theoretical results [3]. Considering learning and exploration rates are decaying, we show that the environment is stabilising (randomness is decreasing along time) and analyse the expected reward and regret of the agents. On this basis, a bound can be defined on the algorithm's expected regret (Theorem 1). Building upon such a bound, we prove that the algorithm is no-regret and converges to an approximate UE (Theorem 2).

THEOREM 1. *The regret achieved by our approach up to time T is bounded by $O\left(\left(\frac{K-1}{TK}\right)\left(\frac{\mu^{T+1}-\mu}{\mu-1}\right)\right)$.*

THEOREM 2. *The algorithm converges to a ϕ-UE, where ϕ is the regret bound of the algorithm.*

Experimental results. Our theoretical results were empirically validated on expanded versions of the Braess graph and compared against standard Q-learning (using reward as reinforcement signal). The obtained regret was indeed within the bound defined in Theorem 1. As compared to standard Q-learning, our approach presented a regret 95% lower, on average. Regarding average travel time, our results were 8% closer to the UE than that of standard Q-learning. Thus, the experiments confirm our theoretical results, showing that our approach is no-regret and approaches the UE.

Next steps. As the next step, I would like to design a novel algorithm that builds upon my theoretical analysis to deliver tighter regret bounds. To this end, I am working on a more efficient exploration mechanism to ensure a sufficient number of samples for each action. I also consider extending our results to the case where agents do not know their routes a priori [2]. This problem cannot be modelled as a stateless MDP. Finally, I also plan to consider mixed strategies.

5. SYSTEM-EFFICIENT EQUILIBRIA

Recall that drivers' selfishness lead to the UE, which is inefficient from a global perspective. The SO, on the other hand, depends on altruistic behaviour, given it is only attainable if some agents take bad routes in favour of global benefit. In general, however, one cannot assume agents are altruist and adhere to such SO-routes: if a better route is available, any rational agent would prefer it instead.

The second front of my thesis aims at bridging the gap between the UE and the SO. Precisely, the idea here is to incorporate some sort of global performance metric into the regret formulation. Consequently, an agent regrets whenever a selfishly taken route burdens the overall performance.

In contrast to previous works, my key contribution lays on developing a method for the agents to estimate such global information. I am investigating how agents can communicate to exchange local perceptions on the system's performance. In other words, whereas an agent does not know the overall average travel time, it can compute a local average based on its peers' performance. As in the literature, the interactions may take place among agents of the same origin (e.g., neighbours) and/or destination (e.g., co-workers).

Challenges and preliminary results. In preliminary tests, we computed an agent's regret using a linear combination of the agent and global travel time (no local estimations were tested up to this point). In the case of pure strategies, we saw that only weak equilibria exist. Consequently, the system may get stuck in a sub-optimal condition. For instance, an agent cannot distinguish between two routes with the same cost, even if one of them negatively affects other agents. To overcome such limitation, mixed strategies could be used so the agent would use each route half of the time.

Next steps. My next plan is to incorporate mixed strategies onto my model and analyse its implications on the current theoretical results. The performance guarantees represent a fundamental aspect of my research. Afterwards, I will develop methods for locally estimating global performance.

Acknowledgments

This work was done in collaboration with Ana L. C. Bazzan (advisor) and Bruno C. da Silva (co-advisor). This research is partially supported by CNPq and CAPES grants.

REFERENCES

[1] A. K. Agogino and K. Tumer. Unifying temporal and structural credit assignment problems. In *AAMAS '04*, pages 980–987, New York, July 2004. IEEE.

[2] B. Awerbuch and R. D. Kleinberg. Adaptive routing with end-to-end feedback: Distributed learning and geometric approaches. In *STOC '04*, pages 45–53, 2004.

[3] G. de O. Ramos, B. C. da Silva, and A. L. C. Bazzan. Learning to minimise regret in route choice. In *AAMAS 2017*, São Paulo, May 2017. IFAAMAS.

[4] K. Waugh, D. Morrill, J. A. Bagnell, and M. Bowling. Solving games with functional regret estimation. In *AAAI '15*, pages 2138–2144. AAAI Press, 2015.