Decreasing Communication Requirements for Agent Specific Rewards in Multiagent Learning

Atil Iscen
Oregon State University
iscena@onid.orst.edu

Chris HolmesParker
Oregon State University
holmespc@onid.orst.edu

Kagan Tumer
Oregon State University
kagan.tumer@oregonstate.edu

ABSTRACT

In many different multiagent domains that require cooperation, success of the agents is heavily dependent on the communication between the agents. For better team performance, shaping individual rewards is essential. As a reward shaping method, difference rewards have shown previous success on many different domains, but the communication requirements are high. This paper defines the set of environment variables on which the agent's reward on the system depends. The definition is used to separate the information needed for the difference reward from the rest of the information about the environment. This concept of the effective area of an agent is explained with an example from the stateless gridworld domain. The experiments show that the performance of the agents with difference reward depends on the amount of information on their effective area. Moreover, if the communication method is designed carefully, the agents can have same quality of reward with 1% to 10% communication.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning

General Terms

Algorithms, Performance

Keywords

Cooperative Learning, Communication, Difference Reward

1. INTRODUCTION

Learning cooperation is a challenging but key problem in many real world applications. To learn cooperation, reward shaping is a highly preferable method to provide better feedback to each individual agent of the system [**?**, **?**]. Difference rewards are reward shaping methods successfully used in many different domains such as air traffic, robot navigation, data routing [**?**, **?**, **?**] . Previous work shows that difference rewards increase the converged behavior and learning time of the agents to a better policy than traditional local and global rewards [**?**]. However, difference rewards are highly dependent on the amount of information and communication about the environment.

In order to collectively optimize system performance, agents need to communicate with each other and share information about the state of the system. However, as the size of the system and the number of agents become increasingly large, the amount of communication and information sharing required quickly becomes problematic. This is especially true when using difference rewards, because of the fact that calculations differ for every agent in the system.

In this paper, the problem stated above is addressed by lowering the communication requirements of the difference reward while keeping same performance. The paper defines the critical set of system variables required for the difference reward. This definition distinguishes the difference between quality and quantity of the information and defines the required communication area to get same performance of the difference reward with less communication.

The remainder of the paper is organized as follows. Section 2 contains the required background knowledge. Section 3 defines the problem of communication requirements in Multiagent Systems. Next, Section 4 introduces stateless gridworld domain that is used both to explain the concepts and to experiment. Section 5 contains the approach used and Section 6 shows the results of the given approach. Section 7 ends the paper with conclusions of the research and future research directions.

2. BACKGROUND

Multiagent Systems (MAS) have successful applications on many real world domains such as air traffic, data routing or robot coordination. Learning in MAS provides benefits such as adaptation to dynamic environments or being more robust to failures . From learning perspective, in a MAS, multiple agents interact by sensing the environment and taking actions. As each agent's action effect the other agents' performances and rewards, the problem has increased complexity than a single agent system. Moreover as the other agents behaviors change over time, the environment is highly dynamic. The agents have to learn to cooperate in addition to learn the domain. Because of these reasons, there are many approaches to modify usual learning methods to MAS [**?**].

From the learning algorithm perspective, although it was developed as a single agent learning algorithm, Reinforcement Learning (RL) is a successful approach to MAL as long as the rewards are set up correctly [**?**]. In RL, the agents learn from their interactions with the environment by sensing and acting [**?**]. The agents and the environment are in a loop where at every timestep t the agent senses the environment with state s_t, takes an action a_t and gets the feedback (reward) of the previous time step r_{t-1}. In this loop, the agents try to learn to take actions that maximizes the feedback that they get.

In some multiagent problems the environment consists of the agents but the agents do not sense the state of the environment. These problems are called stateless problems where the agents take actions and get the reward at every timestep. The goal of the agents is to learn to maximize their reward by learning to adapt to the other agents. Bar problem or congestion domain are good examples to these type of problems [**?**]. These problems do not have a state space problem and provide simpler testbed for learning methods. The general learning rule that is used for the stateless learning agents are: $V(a) \leftarrow (1-\alpha)V(a) + \alpha R$ where $V(a)$ is the value of taking the action a, α is the learning rate and R is the reward of the agent for taking the action a. This paper uses a Stateless Gridworld Problem presented in following sections to both explain and validate the introduced idea.

From type of goal perspective, multiagent problems can be divided into two categories such as Cooperative and Competitive. Cooperative problems are a subfield of multiagent learning problems where the agents try to learn to increase global system (or team) utility by collaborating with each other [**?**]. There are many different approaches to develop cooperative algorithms such as joint learners, game theory and hierarchical learning methods [**?**]. Another method for cooperation is shaping the reward of the individual agents, such that maximizing individual rewards will result in cooperation of the agents [**?**, **?**]. To be able to work on reward shaping methods, next subsections explain different types of rewards and a the difference reward as a successful method for cooperation problems.

2.1 Team Goals and Individual Rewards

Rewards are essential part of the reinforcement learning problems. In cooperative problems, the agents act individually, but their goal is to cooperate and increase the team reward. To provide each agent, there are two trivial types of reward: Global which represents global utility of the team and Local which represents individual effort of the agent itself.

Previous work shows that providing global utility to the agents will not provide the optimal behavior for the learning agents, because an agent can not distinguish the impact of its action on the reward it receives.[**?**].

Another strategy for reward structure is to provide local reward to every individual agent. Local reward is the feedback to the agent depending only on its own action. However, with a local reward, there are no guarantees that the agents actions promote good system behavior.

2.2 Factoredness and Difference Reward

As seen in previous section, the agents that use the global reward (G) and the local reward (L) do not guarantee success for the team. This behavior is explained with two concepts: Factoredness and Learnability. Degree of factoredness of a reward defines the proportion of the individual rewards that are aligned with the global reward. This allows to measure if a different action of the agent results in a better global reward, also results in an increase in the individual reward that it gets. Formally it is defined as:

$$F_{g_i} = \frac{\sum_z \sum_{z'} u\left[(g_i(z) - g_i(z'))(G(z) - G(z'))\right]}{\sum_z \sum_{z'} 1} \quad (1)$$

Where the states z and z' only differ in the state of agent i, and $u[x]$ is the unit step function, equal to 1 if $x > 0$. This definition keeps track of the cases where the change in the individual reward $g_i(z) - g_i(z')$ and the system reward $G(z) - G(z')$ have the same sign. In addition to factoredness, another metric used is learnability. It measures the effect of the agent on the reward. Because it is not in the context of this paper, we omit, but the details can be found in [**?**].

Considering the concepts explained above, the local reward is highly learnable but less factored, and global reward is perfectly factored but not highly learnable. To overcome problems of these two rewards, Difference Reward (D) is a shaped reward that is defined to be more learnable than global reward and more factored than local reward. It is defined as:

$$D_i \equiv G(z) - G(z - z_i + c_i)\ , \quad (2)$$

Where first term $G(z)$ is the global reward of the state z, second term $G(z - z_i + c_i)$ is the global reward of the system where the agent i is taken out and is replaced by an absorbing action. Subtraction of second term from the first term gives the effect of the agent on the system. It is shown to perform better than G and L in many different domains such as [**?**, **?**].

Despite its success, there are some disadvantages of using difference reward: Either the agents or the system should be able to calculate the global reward of the agent. When it is possible for the centralized system to calculate it, it requires recalculation of the global reward without the agent for every agent in the system. If it is calculated by every agent itself, the agents have to observe all the environment and every other agent in the domain or they have to communicate about their actions or states.

3. PROBLEM DEFINITION

As discussed in previous sections, using the difference reward results in better performances. However, each agent requires all the information that is used to calculate global reward and calculates the system reward for each agent. For many different domains, this assumption is not realistic, or requires a lot of calculation. Additionally, even if it is possible, communication is costly, and error prone. It is always a desired behavior to decrease amount of communication.

On the other hand, previous sections explained the difference between performances of difference reward and trivial ones such as global and local. Because of that reason, being able to use similar structure to difference reward even in the imperfect communications is a highly desired solution to multiagent problems. Previous work introduced two different ways of calculating the difference reward in imperfect knowledge of the environment: truncation and estimation [**?**].

Assuming that an agent gets partial information of the system, difference reward can be calculated by using the partial knowledge and ignoring the rest of the system. This approach is called truncation. In contrast, using the partial information to estimate the rest of the system is another approach. If the agent can get the global reward signal in addition to the information, this signal can be used to calculate the first term of the difference reward, which resulted in different approaches to use difference reward in low communication problems [**?**].

In the paper discussed above, the authors give the example of how these different approaches behave in low communication according to global and local rewards. However, we explain the reasons behind these performances of the difference reward approaches in low communication. Additionally, we define the type of information that the difference reward requires, and how one can design a proper way of measuring needed information to expect performances closer to the full communication. To make it easier to explain the concepts, next section defines the stateless gridworld domain that allows to do analysis of different types of communications and different types of the rewards.

4. STATELESS GRIDWORLD DOMAIN

The stateless gridworld domain is a toy multiagent domain based on the cooperation of agents within a gridworld containing points of interests (POI). It is based upon the rover domain in terms of observation and POIs, but the domain is discrete and stateless. Unlike typical gridworld domains, instead of choosing which direction to travel, the agents directly choose a position where they want to be. Although it is more simplified than the usual gridworld domain, it still contains the challenges of a cooperative domain [**?**]. Moreover, the problem given by the domain can be easily associated with real world domain where the agents already have encoded ability to navigate to a chosen point. In this problem the agents try to learn a team behavior to increase the amount of observations made by the team at every time step.

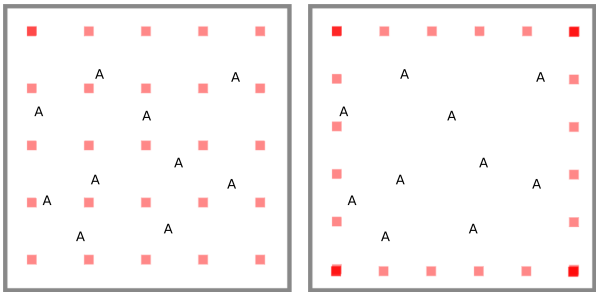
The domain is different from basic stateless problems in terms of dimensionality and existence of the distance metric. Compared to the a simple congestion problem, there exists a distance metric that is naturally defined. The communication restrictions of the agents can be applied either using random sets of agents or using this distance metric. In this domain, we performed experiments with two types of communication restrictions. The first type involved restricted communication rates, which limits the number of agents that any one agent can communicate with at a given time. The second type involved communication-distances, agents were only allowed to communicate with other agents that are within the limit distance.

As expressed above, the main goal for the agents is to observe POIs. Observation of the POIs is defined according to the distance, each POI is observed by the closest agent, and observation value of that POI is determined according to the distance between the POI and the closest agent. For the whole system, team utility that the agents try to increase is the sum of the observations for all of the POIs, defined as:

$$G(z) = \sum_{p \in POIs} \max_{a \in Agents} (0, \beta - min(distance(a, p)) \quad (3)$$

where p represents POI in the system, a represents the agent which minimizes the distance, and β represents the maximum distance that a POI can be observed from.

The system utility for the domain is the global reward calculated by the sum of the observations for the POIs. Local reward for each agent is defined as the amount of observation made by that specific agent. The difference reward discussed above is defined as the system utility with the agent subtracted by the system utility without the agent. When the domain is partially observable due to limited communications, the difference reward used is formed by either truncation of estimation methods discussed in the section 3.



(a) Even POI distribution (b) POIs distributed over the edges

Figure 1: Gridworld types used in the experiments with different POI distributions

For this paper, the number of POIs in the domain is fixed, but to prevent a domain specific approach, the distribution of the POIs is chosen in two different ways. First one distributes the POIs to the gridworld evenly with equal distance between them (i.e. 1 POI at every 10^{th} square). Second approach distributes the POIs over the edges of the gridworld, so that the agents have to learn the distribution over the edges (Figure **??**). There are two main differences in this distribution. First, it makes sure that the agents learn a specific formation other than basic repulsion, second, in low communication cases, in optimal distribution, the agents will not be able to see most of the other agents. Both of these properties make the second distribution problem harder to learn for the agents. Although this paper does not include the cases, the domain is also suitable for the heterogeneous distributions, or POIs with different weights.

5. EFFECTIVE AREA AND DIFFERENCE REWARD

This section contains the main contribution of the paper, the definitions required to explain the quality and the amount of information needed for an agent to calculate difference reward. First, we start by defining effective area of an agent. Effective area of an agent in the system is the set of state variables that the agent's utility on the system depends according to. As an example, if we assume that every cell of the gridworld are the state variables of the environment, the utility of an agent depends on the position of the agent and its surroundings. For example, if we move another agent that is at the other end of the gridworld, this change does not affect the contribution of the agent on the other corner.

For the same example but a different approach, if we assume that environment variables are the positions of the agents, only the variables that represent positions close to the agent are important for the agent. Looking from the other side of the problem, the nature of the difference reward addresses this problem by taking the difference of the system with the agent and without the agent. In this case, assuming that the disappearance of an agent changes the dynamics of a specific region, by subtracting second term from the first term, the region that are not affected by the

agent are already ignored. Combining the definition of the effective area and the nature of the difference reward in the stateless domains, one can conclude that difference reward calculates the contribution of an agent to the system which is limited to some specific environment variables represented by the effective area of that agent at that state.

For example, if the system utility is composed of linear combination of different areas of the system, and a change in some part of the environment only affects certain elements of the sum, the formula $G_z - G_{z-i}$ eliminates the elements in the sum that are not affected by the change. The resulting set of the subtraction represents only the affected elements of the sum, and effective area can be defined as the variables of the system that can can affect these elements (not only the elements, all the variables that can affect these elements).

If we formulate the difference reward for the given gridworld domain, G(z) was defined as a sum of observations of POIs. Assuming that function for measuring the observation for a POI p by an agent a is represented by $f(p, a)$, combining f, Equation **??** and Equation **??** gives:

$$D_i = \sum_{p \in POIs} \max_{a \in Agents} f(p,a) - \sum_{p \in POIs} \max_{a \in Agents_{-i}} f(p,a) \quad (4)$$

The terms of first element and the second element only differ at the places where agent i is the closest agent to the POI p (Case A). So it is 0 in all the other cases.

$$D_i = \sum_{p \in POIs} \begin{cases} \max_{a \in Agents} f(p,a) - \max_{a \in Agents_{-i}} f(p,a) & \text{Case A} \\ 0 & \text{else} \end{cases} \quad (5)$$

For most of the POIs, agent i is out of the range, which gives the ability to cancel elements from both of the terms. Canceling every POI that agent i cannot affect, reduces the Equation **??** to:

$$D_i = \sum_{p \in range(i)} \max_{a \in Agents} f(p,a) - \max_{a \in Agents_{-i}} f(p,a) \quad (6)$$

This reduced definition of the difference reward does not require any information about the far POIs and also the agents that are not in the range of the this small set of POIs. As a description, the resulting information needed is composed of the agents that are in range of the POIs for which the agent i is in the range too. An example to this description is given in Figure **??**.

Considering the stateless gridworld domain, each POI can be observed from a distance of 10, and given the description above, in extreme cases, the effective range can increase up to 20, but most of the cells have a range less than 20. According to the calculations, if an agent has full information about its surrounding within range 20, it can calculate its difference reward without any errors. Moreover, this range does not depend on the size of the gridworld, the same range can be used even for bigger environments.

6. EXPERIMENTS AND RESULTS

The set of experiments that are conducted in this section are ordered as performances of the agents in a specific setup followed by the degree of factoredness and the analysis of the rewards and in this given setup. As described, there are two types of gridworld and two types of communication (Table **??**). All of the experiments contain 20 agents working on a 100×100 gridworld domain. The agents use stateless

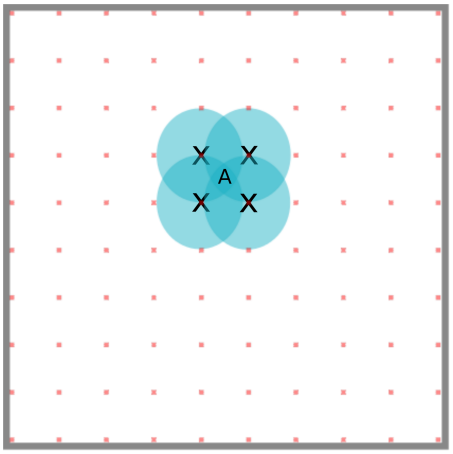


Figure 2: Approximate Effective Area of an agent in gridworld domain. Red dots are the POIs, and the changes in the gridworld outside the radius does not affect the difference reward of the agent

action value learning with learning rate of 0.7 and ϵ-greedy exploration with $\epsilon = 0.9$.

	POI-edge	POI-even
Random	Figures **??**, **??**	Figures **??**, **??**
Distance	Figures **??**, **??**	Figures **??**, **??**

Table 1: Table of the experiments according to the communication type and gridworld type

First experiment is testing the performances of the agents in different levels of communication from 0% to 100%. The communication level allows each agent to only learn (or perceive) the actions (or the positions) of a set of randomly chosen agents. In the 100% communication case, the agents can see all the agents in the gridworld, which leads the truncation or the estimation methods become the actual difference reward. Figure **??** clearly shows that increase in the communication leads to an increase in the performance of the agents. Although this is an expected result, we investigate the reasons by looking at the Figure **??** which shows the factoredness of the reward structure in the defined domain. As the performance graph, we can see an increase at the factoredness with increasing communication level. Both performance and factoredness graphs can be explained with the fact that lower communication levels make the domain less observable and forces the agents to use less information or do an approximation of the complete difference reward.

To make sure that the results hold for non homogeneous distribution, next experiments show the performances of the agents in a gridworld where the POIs are distributed evenly. Figure **??** shows a linear increase in the agents performances. The increase is smaller, because as the POIs are distributed everywhere, the problem is easier for the agents even when they behave randomly. So, the performances in 0 communication is higher than the first experiment, but the same linear increase can be seen. Looking at the factoredness results (Figure **??**), difference reward has the factoredness proportional to the level of communication. On the other hand,

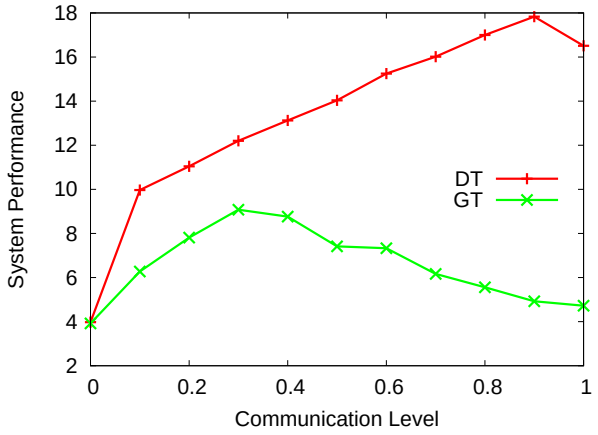


Figure 3: Performances of GT and DT in random communication for POI distribution over the edges. Communicating with random agents increase the performances of the DT, more communication gives better performance.

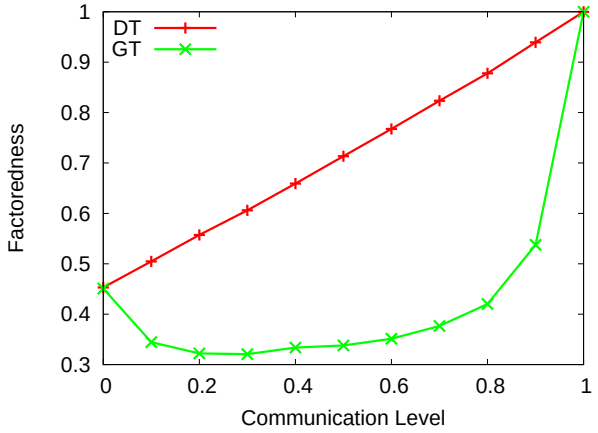


Figure 4: Factoredness of GT and DT in random communication for POI distribution over the edges. As well as performances, communicating with random agents increase the factoredness of DT linearly, more communication gives better factoredness. This is related to more information about effective area

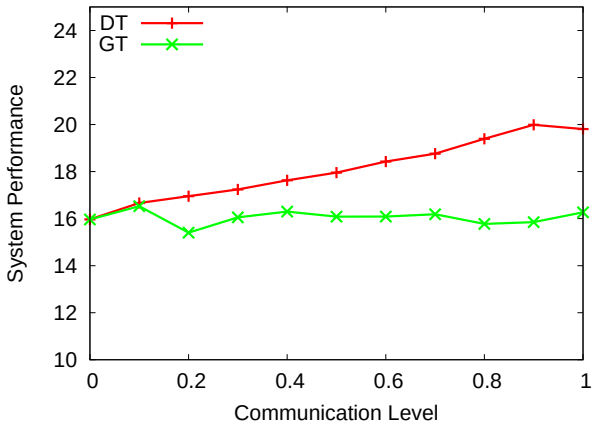


Figure 5: Performances of GT and DT in random communication for evenly distributed POIs. As the problem is easier, performance at 0 is better, but DT have the same increase with more communication.

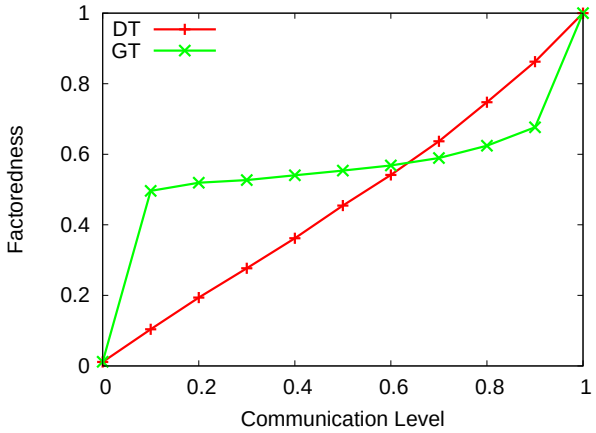


Figure 6: Factoredness of GT and DT in random communication for evenly distributed POIs. Increase in factoredness of DT is exactly linear with respect to communication, because of getting more information about effective area gives more factored problem.

global reward has a monotonically increasing factoredness line, but the shape of the line is different. As this paper is not interested in truncation of global reward, explanation to this factoredness levels are left as a possible future work.

In the 4 results explained so far, the approximately linear increase of the difference reward in both performances and factoredness can be easily explained by effective area of the agent. Communicating with more random agents will increase the chances of getting more information on the effective area. So, linear increase in communication gives linear increase in factoredness and the performances of the agents.

The first four results might be expected, because they conclude with more communication results in better performance. They are also similar to the congestion domain where the communication is defined as acquiring more information about the environment by communication more randomly selected agents. On the other hand, when the agents switch to the limited communication according to the distance with other agents, the critical results can be seen in Figure **??**. The agents show amazing performances at the lower communication levels. One can see that even 10% communication is enough for an agent, to be able to perform close to full communication. Considering the definition of the difference reward where the main disadvantage was stated as full communication requirement, this result lowers this requirement to 10% and makes the difference reward a perfect candidate even for the lower communication cases.

To be able to explain the difference between two types of communications, next experiment **??** shows the factoredness graph for the same experiment. 10% communication level shows the reason of the performance explained before. The truncated difference reward that the agents calculate are

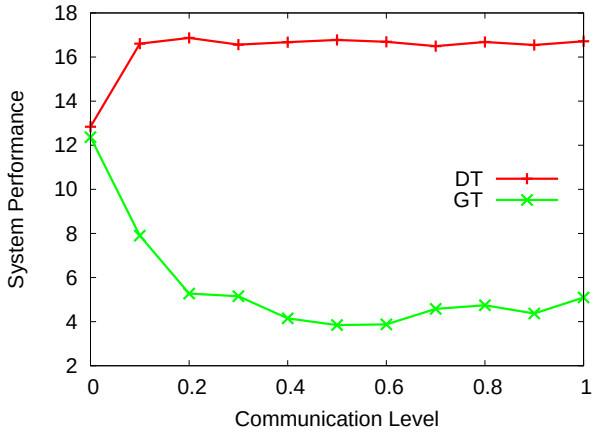


Figure 7: Performances with communication according to distance. Only 10% of communication is enough for DT to perform as well as 100%

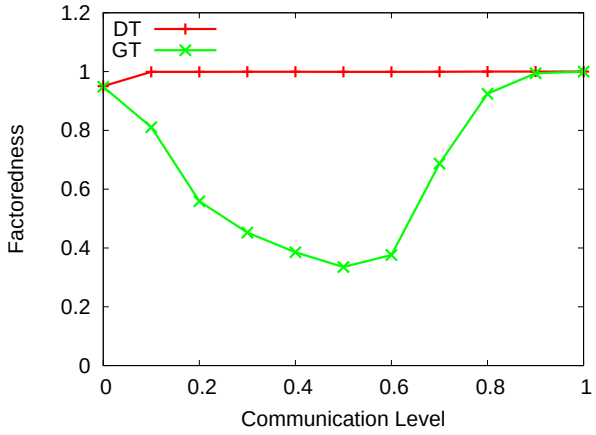


Figure 8: Factoredness with communication according to distance. DT becomes fully factored at 10% communication, because it has most of the information about its effective area.

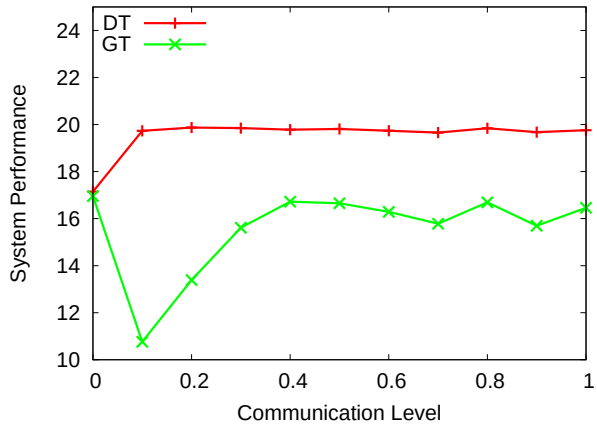


Figure 9: Performances with communication according to distance with evenly distributed POIs. Only 10% of communication is enough for DT to perform as well as 100%

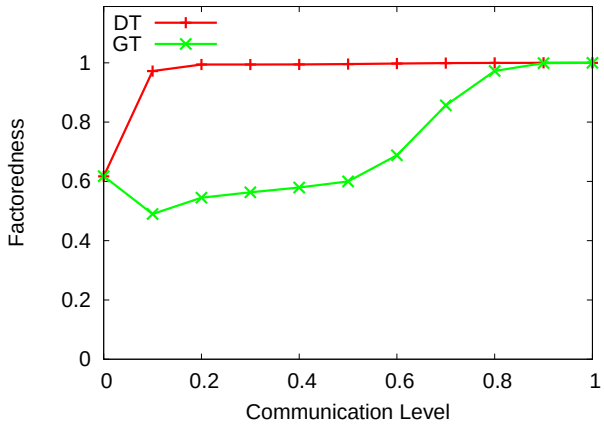


Figure 10: Factoredness with communication according to distance with evenly distributed POIs. DT becomes fully factored at 10% communication

close to totally factored.

The result set for different POI distributions (Figures **??** , **??**) holds the same results giving a critical success within 10% communication. Although having mostly factored reward in a domain where an agent can only see 10% of the world seems unrealistic, it can be easily explained with the effective area of the stateless gridworld domain. Using the description of the effective area, 10% is close enough for an agent to get exact information about its effective area. The information that the agent gets about the rest of the environment is useless for the agent that can use the difference reward to perform at the top level with smaller amount of information.

When the same experiment is repeated for bigger domains (Figure **??**) shows that the results hold for different sizes of the domains and different number of agents. When we increase number of agents and POIs, the effective area of the agent does not depend on the number of agents or size of the domain, it only depends on the limit distance to observe a POI. As there are more agents and more POIs the agents observe more in bigger domains. Moreover as the system gets bigger, the area that 10% communication represents gets bigger, as the size of the effective area is constant, this indicates that the required communication level decreases to below 10%.

7. CONCLUSIONS AND FUTURE WORK

As discussed before, learning coordination is a challenging task. Although the difference reward had a big improvement over the global team reward, problems were arising with less observable, or low communication domains. This paper analyzed the performances of the truncation of the difference reward in a low communication domains, and introduced effective area of an agent to explain the amount and the type of information that the difference reward needs. ‘ As seen in the results, the quality of the information can be measured by information about the effective area, and the information that the agent needs depends on that area instead of all of the environment. As a consequence, we were able to decrease the information needs of the difference reward to

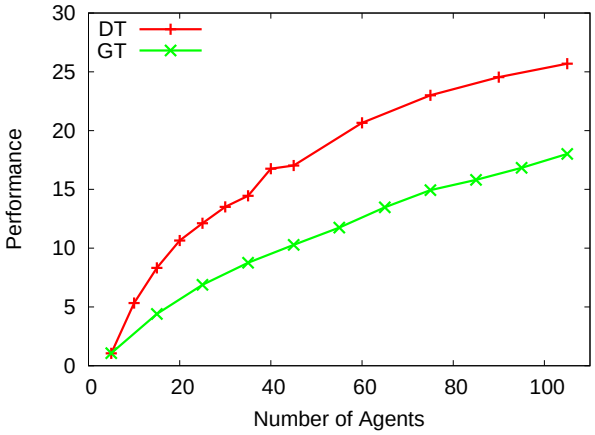


Figure 11: The performances with respect to increasing number of agents. Results hold for bigger systems, as the agents can observe more POIs, performances increase.

10% in an average size the stateless gridworld domain, and the results show that they can perform as good as to the full communication level.

In conclusion, the paper shows that if one can define the effective area of an agent for a specific domain, the agents can benefit the advantages of using the difference reward without suffering the communication requirements. Not only the agents can perform better than global reward, they can also reach full performance of the difference reward with less communication, ignoring the information that the difference reward excludes by its definition.

Future work would extend this into two areas, one future research direction is the automated discovery of the effective states of an agent. Although this can be done by human expertise for some specific domains, automated discovery of the effective range and applications on different types of domains can provide the people basis for same good performance with less communication that does not depend on the size of the environment. Another research opportunity can be approximating the importances of the states in the effective area, and being able to ignore the less important states and having approximately same performance while getting partial information about the effective area of an agent.

8. REFERENCES

[1] A. K. Agogino and K. Tumer. Handling communication restrictions and team formation in congestion games. *Journal of Autonomous Agents and Multi-Agent Systems*, 13(1):97–115, 2006.

[2] A. K. Agogino and K. Tumer. QUICR-learning for multi-agent coordination. In *AAAI*. AAAI Press, 2006.

[3] C. Claus and C. Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the fifteenth national/tenth conference on Artificial intelligence/Innovative applications of artificial intelligence*, AAAI '98/IAAI '98, pages 746–752, Menlo Park, CA, USA, 1998. American Association for Artificial Intelligence.

[4] C. Goldman and S. Zilberstein. Optimizing information exchange in cooperative multi-agent systems. In *Proceedings of the second international joint conference on Autonomous agents and multiagent systems*, pages 137–144. ACM, 2003.

[5] M. Grzes and D. Kudenko. Theoretical and empirical analysis of reward shaping in reinforcement learning. In M. A. Wani, M. M. Kantardzic, V. Palade, L. A. Kurgan, and Y. Qi, editors, *ICMLA; ICMLA*, pages 337–344. IEEE Computer Society, 2009.

[6] S. Kapetanakis and D. Kudenko. Reinforcement learning of coordination in cooperative multi-agent systems. In *Proceedings of the National Conference on Artificial Intelligence*, pages 326–331. Menlo Park, CA; Cambridge, MA; London; AAAI Press; MIT Press; 1999, 2002.

[7] M. Knudson and K. Tumer. Coevolution of heterogeneous multi-robot teams. In *Proceedings of the Genetic and Evolutionary Computation Conference*, Portland, OR, July 2010.

[8] A. D. Laud. *Theory and application of reward shaping in reinforcement learning*. PhD thesis, Champaign, IL, USA, 2004. AAI3130966.

[9] L. Panait and S. Luke. Cooperative multi-agent learning: The state of the art. *Autonomous Agents and Multi-Agent Systems*, 11(3):387–434, 2005.

[10] P. Stone and M. Veloso. Multiagent systems: A survey from a machine learning perspective. *Autonomous Robots*, 8(3):345–383, 2000.

[11] R. Sutton and A. Barto. *Reinforcement learning: An introduction*. The MIT press, 1998.

[12] M. Tan. Multi-agent reinforcement learning: Independent vs. cooperative agents. In *Proceedings of the tenth international conference on machine learning*, volume 337. Citeseer, 1993.

[13] K. Tumer and A. Agogino. Coordinating multi-rover systems: Evaluation functions for dynamic and noisy environments. In *Proceedings of the 2005 conference on Genetic and evolutionary computation*, pages 591–598. ACM, 2005.

[14] K. Tumer and A. Agogino. Distributed agent-based air traffic flow management. In *Proceedings of the Sixth International Joint Conference on Autonomous Agents and Multiagent Systems*, pages 330–337, Honolulu, HI, May 2007.

[15] K. Tumer and A. K. Agogino. A multiagent approach to managing air traffic flow. *Journal of Autonomous Agents and Multi-Agent Systems*, 2010.

[16] S. Williamson, E. Gerding, and N. Jennings. Reward shaping for valuing communications during multi-agent coordination. In *Proceedings of The 8th International Conference on Autonomous Agents and Multiagent Systems-Volume 1*, pages 641–648. International Foundation for Autonomous Agents and Multiagent Systems, 2009.

[17] D. H. Wolpert and K. Tumer. Collective intelligence, data routing and Braess' paradox. *Journal of Artificial Intelligence Research*, 16:359–387, 2002.