Theoretical Considerations of Potential-Based Reward Shaping for Multi-Agent Systems

Sam Devlin
University of York, UK

Daniel Kudenko
University of York, UK

ABSTRACT

Potential-based reward shaping has previously been proven to both be equivalent to Q-table initialisation and guarantee policy invariance in single-agent reinforcement learning. The method has since been used in multi-agent reinforcement learning without consideration of whether the theoretical equivalence and guarantees hold. This paper extends the existing proofs to similar results in multi-agent systems, providing the theoretical background to explain the success of previous empirical studies. Specifically, it is proven that the equivalence to Q-table initialisation remains and the Nash Equilibria of the underlying stochastic game are not modified. Furthermore, we demonstrate empirically that potential-based reward shaping affects exploration and, consequentially, can alter the joint policy converged upon.

Categories and Subject Descriptors

I.2.6 [**Artificial Intelligence**]: Learning; I.2.11 [**Artificial Intelligence**]: Distributed Artificial Intelligence—*Multiagent Systems*

General Terms

Theory, Experimentation

Keywords

Reinforcement Learning, Reward Shaping,
Multiagent Learning, Reward Structures for Learning.

1. INTRODUCTION

Current trends are showing a rise in interest in Multi-Agent Systems (MAS). With multiple, distributed agents a larger set of problem domains can be practically modelled [35]. To control each agent, a reinforcement learning solution can provide adaptive, autonomous, and self-improving agents.

However, whilst reinforcement learning can handle problems with combinatorial state spaces in single-agent problem domains [21, 26], adding more agents to the same environment is a significant challenge [5]. Specifically, as the other agents execute their own actions they subsequently influence the state of the world. This makes the environment appear non-stationary to an individual agent because other agents may concurrently learn and change their behaviour. Unknown to the agent, the same local state-action pair will have a different transition function even though the global state-joint action pair has not changed.

It has been shown in single-agent reinforcement learning that the quicker a learning agent can reach convergence in its policy the more it will benefit from instability in the environment, as it is better suited to adapt to changes [30]. But in MAS the state-action space grows exponentially with the number of agents, which may considerably slow down convergence reducing agents' ability to adapt quickly. Therefore, methods of reducing the time to convergence are of significant importance when implementing reinforcement learning solutions to MAS.

One such method, empirically demonstrated to decrease the time for each individual learning in a common environment to converge on a stable policy, is incorporating heuristic knowledge [17, 25]. However, most existing reinforcement learning algorithms were proposed under the assumption that there is no knowledge available about the problem. This is often not the case; in many practical applications heuristic knowledge can be easily identified by the designer of the system [23], or acquired using reasoning or learning [10].

In single-agent reinforcement learning, potential-based reward shaping has been proven to be a principled and theoretically correct method of incorporating heuristic knowledge into an agent. Provided domain knowledge dependent on states alone, receiving an additional potential-based reward of the correct form does not alter the optimal policy of an agent [20].

To date, applications of potential-based reward shaping to MAS [2, 16] have been studied without published consideration of whether the proofs, originally intended for single-agent problem domains, hold for multi-agent reinforcement learning.

The bulk of our findings, discussed in Section 4, consider the theoretical implications for reward shaping of changing from single-agent problem domains to MAS. This work focuses on the analysis of two fundamental results in single-agent, potential-based reward shaping; the equivalence to Q-table initialisation [33] and the invariance of policies between shaped and non-shaped agents provided [20].

The first remains constant, potential-based reward shaping is equivalent to Q-table initialisation regardless of the

Cite as: Theoretical Considerations of Potential-Based Reward Shaping for Multi-Agent Systems, Sam Devlin and Daniel Kudenko, *Proc. of 10th Int. Conf. on Autonomous Agents and Multiagent Systems (AAMAS 2011)*, Tumer, Yolum, Sonenberg and Stone (eds.), May, 2–6, 2011, Taipei, Taiwan, pp. 225-232.

number of agents learning in the environment. The latter, however, takes new meanings in a MAS. The goal of single-agent reinforcement learning is to compute the policy of maximum reward but with multiple agents, potentially competing, the goal becomes Nash Equilibrium [19]. Therefore, the multi-agent equivalent to policy invariance [20], successfully proven in this paper, is that potential-based reward shaping does not alter the Nash Equilibria of the MAS.

However, potential-based reward shaping can have implications for the joint policy a multi-agent reinforcement learning solution will converge to. As we will show, the final joint policy will still be a Nash Equilibrium of the original system (i.e., before any agents received reward shaping) but may not be the same as the additional reward alters the individual agent's exploration which affects the experiences all agents will have.

We close, in Section 5 by empirically demonstrating our findings but first, to begin, the following section will review existing work and the required background knowledge.

2. EXISTING WORK

2.1 Reinforcement Learning

Reinforcement learning is a paradigm which allows agents to learn by reward and punishment from interactions with the environment [28]. The numeric feedback received from the environment is used to improve the agent's actions. The majority of work in the area of reinforcement learning applies a Markov Decision Process (MDP) as a mathematical model [22].

An MDP is a tuple $\langle S, A, T, R\rangle$, where S is the state space, A is the action space, $T(s,a,s') = Pr(s'|s,a)$ is the probability that action a in state s will lead to state s', and $R(s,a,s')$ is the immediate reward r received when action a taken in state s results in a transition to state s'. The problem of solving an MDP is to find a policy (i.e., mapping from states to actions) which maximises the accumulated reward. When the environment dynamics (transition probabilities and a reward function) are available, this task can be solved using iterative approaches like policy and value iteration [3].

When the environment dynamics are not available, as with most true environments, value iteration cannot be used. However, the concept of an iterative approach remains the backbone of the majority of reinforcement learning algorithms. These algorithms apply so called temporal-difference updates to propagate information about values of states, $V(s)$, or state-action, $Q(s,a)$, pairs [27]. These updates are based on the difference of the two temporally different estimates of a particular state or state-action value. The Q-learning algorithm is such a method [28]. After each transition, $(s,a) \rightarrow (s',r)$, in the environment, it updates state-action values by the formula:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + \gamma \max_{a'} Q(s',a') - Q(s,a)] \quad (1)$$

where α is the rate of learning and γ is the discount factor. It modifies the value of taking action a in state s, when after executing this action the environment returned reward r, and moved to a new state s'.

2.2 Multi-Agent Reinforcement Learning

Applications of reinforcement learning to MAS typically take one of two approaches; multiple individual learners or joint action learners [6]. The former is the deployment of multiple agents each using a single-agent reinforcement learning algorithm. The latter is a group of multi-agent specific algorithms designed to consider the existence of other agents.

Multiple individual learners assume any other agents to be a part of the environment and so, as the others simultaneously learn, the environment appears to be dynamic as the probability of transition when taking action a in state s changes over time. To overcome the appearance of a dynamic environment, joint action learners were developed that extend their value function to consider for each state the value of each possible combination of actions by all agents.

Learning by joint action, however, breaks a common fundamental concept of MAS in which each agent is self motivated and so may not consent to the broadcasting of their action choices. Furthermore, the consideration of the joint action causes an exponential increase in the number of values that must be calculated with each additional agent added to the system. Typically, joint action learning algorithms have only been demonstrated in trivial problem domains [31, 11, 6] whilst applications in complex systems most often implement multiple individual learners [18, 29, 30]. For these reasons, this work will focus on multiple individual learners and not joint action learners. However, these proofs can be extended to cover joint action learners, those we have specifically considered include MiniMax Q-learning [14], Friend-or-Foe Q-learning [15] and Nash-Q [11].

Unlike single-agent reinforcement learning where the goal is to maximise the individual's reward, when multiple self motivated agents are deployed not all agents can always receive their maximum reward. Instead some compromise must be made, typically the system is designed aiming to converge to a Nash Equilibrium [24]. Multiple individual learners will, given sufficient learning time, converge to a point of equilibrium, however, no guarantees can be made that this will be the optimum Nash Equilibrium [6].

To model a MAS, the single-agent MDP becomes inadequate and instead the more general Stochastic Game (SG) is required [5]. A SG of n agents is a tuple $\langle S, A_1, ..., A_n, T, R_1, ..., R_n\rangle$, where S is the state space, A_i is the action space of agent i, $T(s,A,s') = Pr(s'|s,A)$ is the probability that joint action A in state s will lead to state s', and $R_i(s,a,s')$ is the immediate reward r received by agent i when action a taken in state s results in a transition to state s' [9].

3. REWARD SHAPING

The immediate reward r, which is in the update rule given by Equation 1, represents the feedback from the environment. The idea of *reward shaping* is to provide an additional reward which will improve the convergence of the learning agent with regard to the learning speed [20, 23]. This concept can be represented by the following formula for the Q-learning algorithm:

$$Q(s,a) \leftarrow Q(s,a) + \alpha[r + F(s,s') + \gamma \max_{a'} Q(s',a') - Q(s,a)] \quad (2)$$

where $F(s,s')$ is the general form of the shaping reward.

Even though reward shaping has been powerful in many

experiments it quickly became apparent that, when used improperly, it can change the optimal policy [23]. To deal with such problems, potential-based reward shaping was proposed [20] as the difference of some potential function Φ defined over a source s and a destination state s':

$$F(s, s') = \gamma\Phi(s') - \Phi(s) \quad (3)$$

where γ must be the same discount factor as used in the agent's update rule (see Equation 1).

Ng et al. [20] proved that potential-based reward shaping, defined according to Equation 3, guarantees learning a policy which is equivalent to the one learnt without reward shaping in both infinite and finite horizon MDPs.

Wiewiora [33] later proved that an agent learning with potential-based reward shaping and no knowledge-based Q-table initialisation will behave identically to an agent without reward shaping when the latter agent's value function is initialised with the same heuristic knowledge represented by $\Phi(s)$. This is an important fact, because when function approximation is used in big environments, where the structural properties of the state space are not clear, it is not easy to initialise the value function. Potential-based reward shaping represents a flexible and theoretically correct method to incorporate background knowledge regarding states into reinforcement learning algorithms.

3.1 Reward Shaping In Multi-Agent Systems

Incorporating heuristic knowledge has been shown to be beneficial in multi-agent reinforcement learning [2, 16, 17, 25]. However, some of the previous examples did not use potential-based functions to shape the reward [17, 25] and could potentially, therefore, suffer from introducing beneficial cyclic policies that cause convergence to an unintended behaviour as demonstrated previously in a single-agent problem domain [23].

The remaining applications that were potential-based [2, 16], demonstrated an increased probability of convergence to a higher value Nash Equilibrium. As it has long been established that multiple individual learners are not guaranteed to converge to the optimal Nash Equilibrium [6], a number of methods to increase the probability of this occurring have already been devised. Amongst them are COIN [34] and myopic heuristics [6]. However, these methods require knowledge of the reward function or the joint action. Potential-based reward shaping can similarly increase the probability of convergence to the optimal Nash Equilibrium provided a good heuristic, but does so without requiring either of these specific pieces of knowledge which are commonly unavailable in MAS applications.

Both applications of potential-based reward shaping were published with no consideration of whether the proofs of guaranteed policy invariance hold in multi-agent reinforcement learning or how they affect the joint policy at time of convergence. Starting in the following section, our contribution fills this gap in knowledge and provides the theoretical results to explain these previous empirical studies.

4. THEORY

To discuss the implications of using potential-based reward shaping in MAS we must consider the differences between single-agent and multi-agent reinforcement learning. SGs, unlike MDPs, share amongst all agents a common transition function and common states but neither of these are affected by shaping the reward function of one or more of the agents. Although the agents may change their own policy and alter their exploration path due to the additional potential-based reward, this does not change the dynamics (transition function or states) of the environment, nor the set of actions the agent can take.

In fact the only elements of a SG to change when one or more agent implements potential-based reward shaping are the individual reward functions of those agents. If, as we will later show to be true in Section 4.2, these alterations to the individual reward functions do not change the best response policy of a shaped agent given a fixed set of policies followed by all other agents, the Nash Equilibria of the underlying SG remain constant regardless of how many agents are using potential-based reward shaping.

Formally, this argument will be completed by showing, in the following sub-section, that potential-based reward shaping in MAS is equivalent to Q-table initialisation and then, in Section 4.2, that it does not alter the Nash Equilibria of the MAS. Both of these findings, as we will discuss in Section 4.3, has implications for the eventual policy that will be converged upon.

4.1 Potential-Based Reward Shaping And Q-Value Initialisation Are Equivalent

The proof of Wiewiora [33] of the equivalence of potential-based reward shaping and Q-value initialisation was published in the context of single agent problem domains but also holds for problem domains with multiple individual learners.

From [33] we quote:

> *Theorem 1* Given the same sequence of experiences during learning, $\Delta Q(s, a)$ always equals $\Delta Q'(s, a)$.

where $Q(s, a)$ is the modelled value function of an agent learning with potential-based reward shaping and $Q'(s, a)$ is the modelled value function of an agent learning with Q-value initialisation.

The original proof uses a fixed sequence of experiences for both agents. The theory can be extended to multiple individual learners simply by extending the definition of the sequence experienced from the 4-tuple $\langle s, a, r, s' \rangle$ to the $2n + 2$-tuple $\langle s, a_1, a_2, ..., a_n, r_1, r_2, ..., r_n, s' \rangle$. Using the extended sequence and the inductive proof from [33] the following proves that Theorem 1 holds also for multi-agent reinforcement learning.

Proof By Induction

Consider any arbitrary agent i from the set of all agents. As before, $Q(s, a)$ is the modelled value function when the agent is learning with potential-based reward shaping and $Q'(s, a)$ is the modelled value function had the same agent learnt without reward shaping but with Q-value initialisation. The former agent will later be referred to as L and the latter as L'.

Agent L will update its Q-values by the rule:

$$\begin{aligned} Q_i(s, a) \quad \leftarrow \quad & Q_i(s, a) + \\ & \underbrace{\alpha \left(r_i + F(s, s') + \gamma \max_{a'} Q_i(s', a') - Q_i(s, a) \right)}_{\delta Q_i(s,a)} \end{aligned} \quad (4)$$

where $F(s, s')$ is the potential-based reward shaping function and $\delta Q_i(s, a)$ is the amount (scaled by α) that the Q value will be updated by. The current Q-values of Agent L can be represented formally as the initial value plus the change since:

$$Q_i(s, a) = Q_i^0(s, a) + \Delta Q_i(s, a) \tag{5}$$

where $Q_i^0(s, a)$ is agent i's initial Q-value of state-action pair (s, a). Similarly agent L' updates its Q-values by the rule:

$$Q_i'(s, a) \leftarrow Q_i'(s, a) + \underbrace{\alpha\left(r_i + \gamma \max_{a'} Q_i'(s', a') - Q_i'(s, a)\right)}_{\delta Q_i'(s,a)} \tag{6}$$

And its current Q-values can be represented formally as:

$$Q_i'(s, a) = Q_i^0(s, a) + \Phi(s) + \Delta Q_i'(s, a) \tag{7}$$

where $\Phi(s)$ is the potential for state s.

Base Case

Before either agent experiences anything, the Q-tables of L and L' are both their respective initial values, and therefore both ΔQ_i and $\Delta Q_i'$ are uniformly zero.

Inductive Case

Assuming $\Delta Q_i = \Delta Q_i'$, both L and L' will be updated by the same amount in response to experience $\langle s, a_1, a_2, ..., a_n, r_1, r_2, ..., r_n, s' \rangle$. First consider the update performed by L:

$$\begin{aligned}
\delta Q_i(s, a) &= r_i + F(s, s') + \gamma \max_{a'} Q_i(s', a') - Q_i(s, a) \\
&= r_i + \gamma\Phi(s') - \Phi(s) \\
&\quad + \gamma \max_{a'}(Q_i^0(s', a') + \Delta Q_i(s', a')) \\
&\quad - Q_i^0(s, a) - \Delta Q_i(s, a)
\end{aligned} \tag{8}$$

Now consider the update performed by L':

$$\begin{aligned}
\delta Q_i'(s, a) &= r_i + \gamma \max_{a'} Q_i'(s', a') - Q_i'(s, a) \\
&= r_i + \gamma \max_{a'}(Q_i^0(s', a') + \Phi(s') + \Delta Q'(s'a')) \\
&\quad - Q_i^0(s, a) - \Phi(s) - \Delta Q'(s, a) \\
&= r_i + \gamma \max_{a'}(Q_i^0(s', a') + \Phi(s') + \Delta Q(s'a')) \\
&\quad - Q_i^0(s, a) - \Phi(s) - \Delta Q(s, a) \\
&= r_i + \gamma\Phi(s') - \Phi(s) \\
&\quad + \gamma \max_{a'}(Q_i^0(s', a') + \Delta Q_i(s', a')) \\
&\quad - Q_i^0(s, a) - \Delta Q_i(s, a) \\
&= \delta Q_i(s, a)
\end{aligned} \tag{9}$$

Therefore, the Q-tables of both L and L' are both updated by the same value and so ΔQ_i and $\Delta Q_i'$ remain equal. □

Given that Theorem 1 of [33] holds for the multi-agent context then so too does Theorem 2, again quoted from [33]:

> *Theorem 2* If L and L' have learnt on the same sequence of experiences and use an advantage-based policy, they will have an identical probability distribution for their next action.

where an advantage-based policy is one that chooses actions based not on the absolute magnitude of the Q-values but on their relative differences within the current state. Examples of advantage-based policies include greedy, ϵ-greedy and Boltzmann soft-max.

This is immediately apparent when considering both $\Delta Q_i = \Delta Q_i'$ from Theorem 1 and Equations 5 and 7. As the difference between the Q-values of agent L and agent L' are the potential of the state, the difference is consistent across all actions in any given state. Therefore, the actions maintain the same relative differences allowing an advantage-based policy to make the same action decisions.

Effectively, at any time in learning L and L' will behave the same way (make the same decisions with the same probabilities). To conclude, whether an agent is shaped or initialised it will have the same effect on all other agents in the environment, the learning dynamics are not changed by using one method or the other and the agents as a collective whole will converge or not upon the same joint policy regardless of whether the agent was shaped or initialised.

Finally, although the proof here was written specifically for Q-learning, this was simply in keeping with the original work of [33]. In single-agent problem domains the equivalence of Q-table initialisation and potential-based reward shaping can be proven also in SARSA and other temporal difference algorithms [33]. Similar extensions to multi-agent, as above, are possible also for these extensions.

4.2 Potential-Based Reward Shaping Does Not Alter The Nash Equilibria Of A Stochastic Game

As already established the common goal of MARL is a Nash Equilibrium. The typical concern of modifying a reward function is that the original goals of the agent will be altered. Ng showed previously that in the single-agent context, the optimum policy was unchanged by the introduction of reward shaping provided the function was potential-based [20]. To extend this to MARL we must now consider whether implementing the same reward shaping in one or more agents in a SG will alter its points of equilibrium.

Formally a Nash Equilibrium in a SG is:

$$\forall i \in 1 \ldots n, \pi_i \in \Pi_i | R_i(\pi_i^{NE} \cup \pi_{-i}^{NE}) \geq R_i(\pi_i \cup \pi_{-i}^{NE}) \tag{10}$$

where n is the number of agents, Π_i is the set of all possible policies of agent i, R_i is the reward function for agent i, π_i^{NE} is a specific policy of agent i and π_{-i}^{NE} is the joint policy of all agents except agent i following their own fixed specific policy. If the inequality holds for all agents, the joint policy of each agent following its policy π_i^{NE} is a Nash Equilibrium.

Now consider any arbitrary agent i from the set of all agents. For the inequality above to hold for agent i, we must consider the set Π_i^{NE} of all joint policies consisting of each possible policy of agent i combined with π_{-i}^{NE}. Formally, this set contains:

$$\forall \pi_i \in \Pi_i | (\pi_i \cup \pi_{-i}^{NE}) \tag{11}$$

Each fixed joint policy in the set Π_i^{NE} will generate a fixed infinite sequence of experiences when followed consistently from the current state s_0 of the form:

$$\begin{aligned}\bar{s} \quad = \quad & s_0, a_{0,0}, a_{0,1}, \ldots, a_{0,n}, r_{0,0}, r_{0,1}, \ldots, r_{0,n}, \ldots, \\ & s_\infty, a_{\infty,0}, a_{\infty,1}, \ldots, a_{\infty,n}, r_{\infty,0}, r_{\infty,1}, \ldots, r_{\infty,n}, \ldots\end{aligned} \tag{12}$$

where s_j is the state at time j, $a_{j,i}$ is the action taken by agent i at time j and $r_{j,i}$ is the reward received by agent i at time j.

Then using the proof of [1], we can show the difference of the return received by agent i when following any arbitrary fixed sequence with or without potential-based reward shaping is the potential of the state s_0.

Proof

The return for agent i when experiencing sequence $\bar{s}$ in a discounted framework without shaping is:

$$U_i(\bar{s}) = \sum_{j=0}^{\infty} \gamma^j r_{j,i} \tag{13}$$

Now consider the same agent but with a reward function modified by adding a potential-based reward function. The return of the shaped agent experiencing the same sequence $\bar{s}$ is:

$$\begin{aligned} U_{i,\Phi}(\bar{s}) &= \sum_{j=0}^{\infty} \gamma^j (r_{j,i} + F(s_j, s_{j+1})) \\ &= \sum_{j=0}^{\infty} \gamma^j (r_{j,i} + \gamma\Phi(s_{j+1}) - \Phi(s_j)) \\ &= \sum_{j=0}^{\infty} \gamma^j r_{j,i} + \sum_{j=0}^{\infty} \gamma^{j+1}\Phi(s_{j+1}) - \sum_{j=0}^{\infty} \gamma^j \Phi(s_j) \\ &= U_i(\bar{s}) + \sum_{j=1}^{\infty} \gamma^j \Phi(s_j) - \Phi(s_0) - \sum_{j=1}^{\infty} \gamma^j \Phi(s_j) \\ &= U_i(\bar{s}) - \Phi(s_0) \end{aligned} \tag{14}$$

□

Therefore, any policy that previously maintained the inequality of Equation 10 will still maintain the inequality. Formally, and more strictly we can conclude:

$$\begin{aligned}\forall \pi_i \in \Pi_i | \quad & (R_i(\pi_i^{NE} \cup \pi_{-i}^{NE}) \geq R_i(\pi_i \cup \pi_{-i}^{NE})) \leftrightarrow \\ & (R_{i,\Phi}(\pi_i^{NE} \cup \pi_{-i}^{NE}) \geq R_{i,\Phi}(\pi_i \cup \pi_{-i}^{NE}))\end{aligned} \tag{15}$$

where $R_{i,\Phi}$ is the reward function of agent i when receiving both the environmental reward and the potential-based reward shaping.

As implementing reward shaping only affects the reward function of that agent, the remaining agents will also still maintain the same policies as part of the Nash Equilibria. Whether the group will converge to this point depends on the learning algorithm used and is outside of this proof. However, it suffices to say that regardless of how many agents in the MAS are or are not implementing potential-based reward shaping the points of equilibrium will remain constant.

4.3 Potential-Based Reward Shaping Alters Exploration

In Section 4.1 we showed that an agent in a MAS receiving potential-based reward shaping is equivalent to one whose Q-table was initialised with each state s set to the potential $\Phi(s)$ of that state. However, the implications of this proof in a MAS extend past showing that two methods of introducing domain knowledge are equivalent. Instead, it is worth considering the results of [32], in which Wellman and Hu showed that the joint policy converged upon in a learning MAS was highly sensitive to initial belief. This clearly applies directly to Q-table initialisation, where the initial values directly represent some initial belief, and therefore, given that we have shown the equivalence between initialisation and shaping, also applies to potential-based reward shaping. This can be reasoned intuitively by considering the following.

The MDP of an agent deployed in a common environment with other learning agents does not hold the Markov property as the transition probabilities are subject to change with the unseen but changing policies of the other agents. Therefore, the convergence to optimal policy guarantees of Q-learning do not hold. This has been demonstrated empirically in multi-agent reinforcement applications with multiple Q-learners converging to sub-optimal joint policies [2].

Shaping alters the path of exploration an agent takes. In single-agent reinforcement learning, as convergence to the optimal policy is guaranteed, this only affects the time taken to reach convergence. If a good heuristic, is used the time will be reduced as the number of sub-optimal actions taken will be reduced, but similarly if a bad heuristic is used the agent will take longer to converge to the optimal policy.

The concept of an optimal policy in MAS is not as clear. We have identified Nash Equilibrium as the typical goal of multi-agent reinforcement learning, but this does not necessarily identify a single goal. Most applications, with the exception of the very trivial, will have multiple points of equilibrium. Multiple individual learners will converge to one of these equilibrium, but whether it will be the optimum cannot be guaranteed [6]

With multiple agents in the same environment, altering the exploration of one will change the experiences of all agents [12, 13]. The change in actions chosen by even just one agent now receiving potential-based reward shaping will result in different state transitions. The agents will then explore different areas of the joint policy space and, with multiple points of equilibrium possible, may converge to a different equilibrium then had the agent not received the reward shaping and subsequently not have altered its individual exploration path.

Therefore, in multi-agent problem domains, without the guarantee of convergence to a single optimum goal, shaping can lead to convergence on a different joint policy. This was empirically demonstrated by Babes and Littman [2], where a shaped agent was able to lead a non-shaped agent to convergence on a joint policy of higher average reward. When shaping one or more agents in an environment with multiple learning agents, a good heuristic will encourage higher global utility similar to how in single-agent problem domains the use was preferably to reduce the time taken to converge. Unfortunately, the techniques can also have a detrimental effect encouraging miscoordination and/or lead the agents to converge on a less beneficial joint policy by directing the

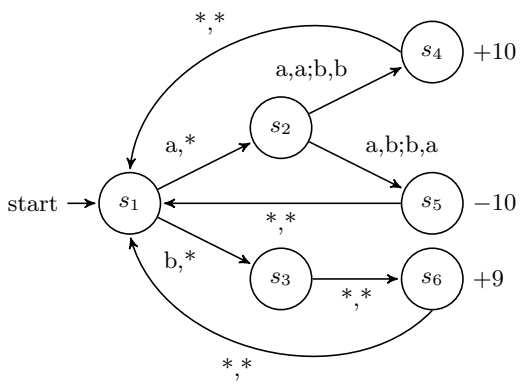

Figure 1: Boutilier's Coordination Game

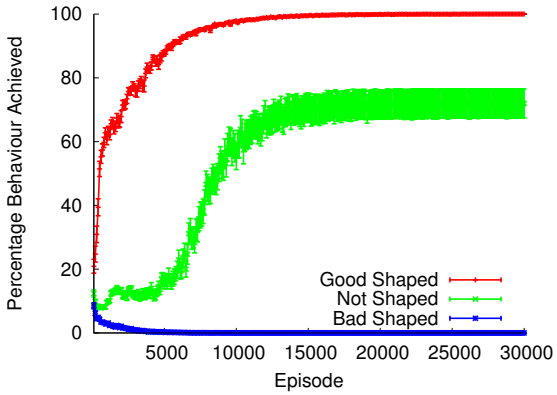

Figure 2: Optimal Nash Equilibrium

agents away from frequently, or possibly ever, experiencing the equilibrium reached by non-shaped agents and instead trapping them in a sub-optimal point of equilibrium.

To support and illustrate these claims the following section will present an empirical study that is typically characteristic of implementing potential-based reward shaping in a MAS.

5. EMPIRICAL DEMONSTRATION

To demonstrate the theorised effects of potential-based reward shaping, an empirical study of a game based on Boutilier's coordination game [4] will be presented here.

The game, illustrated in Figure 1, has six stages and two agents, each capable of two actions (a or b). The first agent's first action choice in each episode decides if the agents will move to a state guaranteed to reward them minimally (s_3) or to a state where they must co-ordinate to receive the highest reward (s_2). However, in state s_2 the agents are at risk of receiving a large negative reward if they do not choose the same action.

In Figure 1, each transition is labelled with one or more action pairs such that the pair $a, *$ means this transition occurs if agent 1 chooses action a and agent 2 chooses either action. When multiple action pairs result in the same transition the pairs are separated by a semicolon(;).

The game has three joint policy Nash Equilibria; the joint policy of opting for the safety state s_3 or the two joint policies of moving to state s_2 and coordinating on both choosing a or b. Any joint policy receiving the negative reward is not a Nash Equilibrium, as the first agent can choose to change its first action choice and so receive a higher reward by instead reaching state s_3.

Three sets of agents will be the focus of these experiments. All agents, in all sets, will learn by Q-learning with an ϵ-greedy policy and discount factor (γ) of 1. One set will receive no reward shaping, to illustrate the average performance without heuristic knowledge, another set will receive potential-based reward shaping from a good heuristic whilst the final set receives shaping from a poor heuristic.

The good heuristic, designed to encourage co-operation, gives states s_1, s_2 and s_4 the potentials 5, 10 and 15 respectively. All other states receive a potential of 0. Therefore, any transition from states s_1 to s_2 or s_2 to s_4 will receive an additional reward of +5 but transitioning instead from state s_2 to s_5 will receive an additional reward of −10. Alternatively, the bad heuristic is designed to encourage miscoordination and so potentials of 5, 10 and 15 are given instead to states s_1, s_2 and s_5 respectively. Again all other states receive a potential of 0.

Our experimental results are intended to show, provided a good heuristic, the increased probability of converging to the joint policies of higher global utility (those achieving coordination in state s_2). Alternatively, provided a bad heuristic, the agents will demonstrate that the Nash Equilibria have not changed and so converge still to one of the three original joint policy Nash Equilibria.

5.1 Results

All experiments were run for 100,000 episodes (300,000 action choices) and repeated 100 times. The results, illustrated in Figures 2, 3 and 4, plot the mean percentage of the last 100 episodes performing the optimal, safety and sub-optimal joint policies respectively. All figures include error bars illustrating the standard error from the mean. For clarity, graphs are plotted only up to 30,000 episodes as by this time all experiments had converged to a stable joint policy.

Figure 2 shows that, for this relatively simple game, multiple individual learners alone can only converge to the optimal behaviour 72% of the time. Whereas, provided a good heuristic, potential-based reward shaping can increase the probability of convergence to this Nash Equilibrium to 100%.

As theorised, provided a bad heuristic, the effect on the global utility can be detrimental. The probability of achieving optimal behaviour, with a potential function encouraging miscoordination, rapidly drops and converges on 0%.

Instead, as illustrated by Figure 3, the poorly shaped agents converge to the safety Nash Equilibrium. Despite the miscoordination state (s_5) receiving the largest potential, the agents do not converge to the sub-optimal behaviour, as illustrated by Figure 4.

Figure 4, highlights that agents with no shaping or potential-based reward shaping never converge to consistently perform the sub-optimal joint policy. This is because miscoordination in this game is not a Nash Equilibrium, both with and without potential-based reward shaping. Regardless of which joint policy is encouraged, if the additional reward is potential based, the Nash Equilibria remain constant.

However, Figure 4 illustrates the behaviour of an addi-

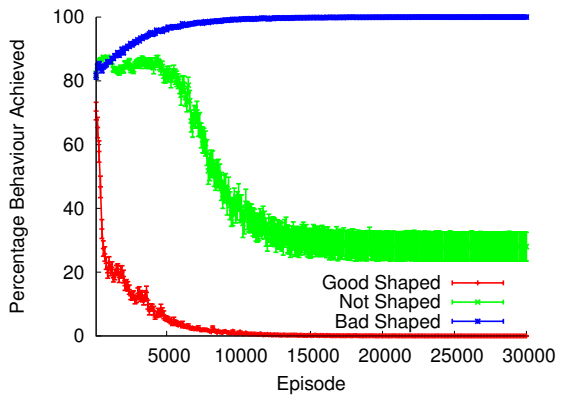

Figure 3: Safety Nash Equilibrium

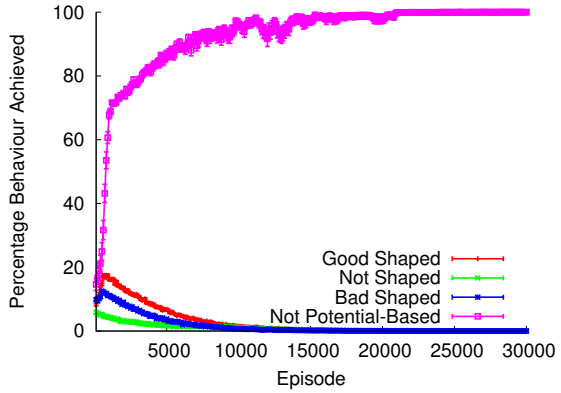

Figure 4: Sub-Optimal Behaviour

tional set of agents. These agents receive an additional reward on state transitions that is not potential based. Specifically they are rewarded 5, 10 and 30 upon entering states s_1, s_2 and s_5 respectively. These agents converge to a joint policy representative of the sub-optimal behaviour. This has occurred because if additional rewards are not potential-based they can change the Nash Equilibria of a SG.

Finally, the learning performance of both sets of shaped agents favourably supports the use of potential-based reward shaping in MAS. The poorly shaped agents converge to the safety Nash Equilibrium after just 10,000 episodes whilst without shaping it takes agents 29,000 episodes to converge. More significantly, after only 2000 episodes agents receiving reward shaping from a good heuristic are more likely to achieve the optimal Nash Equilibrium than non-shaped agents ever will.

6. CONCLUSION

In conclusion, this paper shows how two fundamental papers in single-agent reward shaping [20, 33] can be extended to provide similar guarantees in multi-agent reinforcement learning.

Specifically, we have proven that a potential-based shaped agent is still equivalent to an agent with initial Q-values set to the potential of each state regardless of how many exist within the same environment.

Furthermore, we have also proven that rewarding any number of agents within a MAS with additional potential-based rewards has no subsequent effect on the Nash Equilibria of the underlying SG.

Potential-based reward shaping affects the exploration of the shaped agent. Therefore, it can change the joint policy converged upon as even just one agent's modified exploration can sufficiently redirect the search of joint policy space to converge to a different point of equilibrium.

Although the agents may now converge to a different joint policy, the latter of the two proofs guarantees that the new joint policy was also a goal of the unshaped agents.

Whether the goal achieved is the Nash Equilibrium of highest global utility, is dependent on the agents' learning algorithms. With multiple individual learners, no guarantee of convergence to the highest utility Nash Equilibrium is provided. However, potential-based reward shaping can, dependent on the heuristic, either increase or decrease the probability of converging to equilibria of higher global utility as demonstrated in our empirical study.

Given a joint action learner guaranteed under fixed conditions to converge, such as NashQ [11], it is possible to construct similar proofs as those shown here. Agents learning by joint action and receiving potential-based reward shaping benefit from consistent Nash Equilibria, modified exploration to decrease the number of sub-optimal action decisions and guaranteed convergence.

It is also the authors' expectation that potential-based advice [8], an extension of potential-based reward shaping to include heuristics based on actions as well as states, could similarly be extended to guarantee consistent Nash Equilibria when applied to multi-agent reinforcement learning. Recent empirical work supports these expectations [7].

The work here has been based entirely in fully observable problem domains, which some may consider uncharacteristic of MAS. However, by shaping agents based on the potential of observations (as opposed to fully observed states) the same arguments and proofs can be used to show similar theoretical expectations in partially observable problem domains. Namely, the Nash Equilibria of a partially observable problem domain would remain the same but the agents exploration will alter and so convergence may be to a different point of equilibrium or, given an unsuitable heuristic, may not converge at all.

In closing, adding potential-based reward shaping to multiple individual learners does not alter the Nash Equilibria but can, provided suitable heuristics, increase the probability of convergence to a higher global utility and decrease the time to convergence.

7. ACKNOWLEDGEMENTS

We would like to thank M.Babes, M.Littman and M.Grześ for their input during the development of these ideas and C.Poskitt for his time spent proof reading.

8. REFERENCES

[1] J. Asmuth, M. Littman, and R. Zinkov. Potential-based shaping in model-based reinforcement learning. In *Proceedings of the Twenty-Third AAAI Conference on Artificial Intelligence*, pages 604–609, 2008.

[2] M. Babes, E. de Cote, and M. Littman. Social reward shaping in the prisoner's dilemma. In *Proceedings of*

the 7th International Joint Conference on Autonomous Agents and Multiagent Systems, volume 3, pages 1389–1392, 2008.

[3] D. P. Bertsekas. *Dynamic Programming and Optimal Control (2 Vol Set)*. Athena Scientific, 3rd edition, 2007.

[4] C. Boutilier. Sequential optimality and coordination in multiagent systems. In *International Joint Conference on Artificial Intelligence*, volume 16, pages 478–485. Citeseer, 1999.

[5] L. Busoniu, R. Babuska, and B. De Schutter. A Comprehensive Survey of MultiAgent Reinforcement Learning. *IEEE Transactions on Systems Man & Cybernetics Part C Applications and Reviews*, 38(2):156, 2008.

[6] C. Claus and C. Boutilier. The dynamics of reinforcement learning in cooperative multiagent systems. In *Proceedings of the National Conference on Artificial Intelligence*, pages 746–752, 1998.

[7] S. Devlin, M. Grześ, and D. Kudenko. Multi-agent, potential-based reward shaping for RoboCup KeepAway. In *Proceedings of The Tenth Annual International Conference on Autonomous Agents and Multiagent Systems (AAMAS)*, 2011.

[8] G. C. Eric Wiewiora and C. Elkan. Principled methods for advising reinforcement learning agents. In *Proceedings of the Twentieth International Conference on Machine Learning*, 2003.

[9] J. Filar and K. Vrieze. *Competitive Markov decision processes*. Springer Verlag, 1997.

[10] M. Grześ and D. Kudenko. Plan-based reward shaping for reinforcement learning. In *Proceedings of the 4th IEEE International Conference on Intelligent Systems (IS'08)*, pages 22–29. IEEE, 2008.

[11] J. Hu and M. Wellman. Nash Q-learning for general-sum stochastic games. *The Journal of Machine Learning Research*, 4:1039–1069, 2003.

[12] 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.

[13] S. Kapetanakis and D. Kudenko. Reinforcement learning of coordination in heterogeneous cooperative multi-agent systems. pages 119–131, 2004.

[14] M. Littman. Markov games as a framework for multi-agent reinforcement learning. In *Proceedings of the eleventh international conference on machine learning*, volume 157, page 163. Citeseer, 1994.

[15] M. Littman. Friend-or-foe Q-learning in general-sum games. In *Machine Learning - International Workshop then Conference*, pages 322–328, 2001.

[16] B. Marthi. Automatic shaping and decomposition of reward functions. In *Proceedings of the 24th International Conference on Machine learning*, page 608. ACM, 2007.

[17] M. Matarić. Reinforcement learning in the multi-robot domain. *Autonomous Robots*, 4(1):73–83, 1997.

[18] M. Mihaylov, K. Tuyls, and A. Nowé. Decentralized Learning in Wireless Sensor Networks. *Adaptive and Learning Agents*, pages 60–73, 2009.

[19] J. Nash. Non-cooperative games. *Annals of mathematics*, 54(2):286–295, 1951.

[20] A. Y. Ng, D. Harada, and S. J. Russell. Policy invariance under reward transformations: Theory and application to reward shaping. In *Proceedings of the 16th International Conference on Machine Learning*, pages 278–287, 1999.

[21] J. Peters, S. Vijayakumar, and S. Schaal. Reinforcement learning for humanoid robotics. In *Proceedings of Humanoids2003, Third IEEE-RAS International Conference on Humanoid Robots*, 2003.

[22] M. L. Puterman. *Markov Decision Processes: Discrete Stochastic Dynamic Programming*. John Wiley & Sons, Inc., New York, NY, USA, 1994.

[23] J. Randløv and P. Alstrom. Learning to drive a bicycle using reinforcement learning and shaping. In *Proceedings of the 15th International Conference on Machine Learning*, pages 463–471, 1998.

[24] Y. Shoham, R. Powers, and T. Grenager. If multi-agent learning is the answer, what is the question? *Artificial Intelligence*, 171(7):365–377, 2007.

[25] P. Stone and M. Veloso. Team-partitioned, opaque-transition reinforcement learning. In *Proceedings of the third annual conference on Autonomous Agents*, pages 206–212. ACM, 1999.

[26] R. Sutton. Generalization in Reinforcement Learning: Successful Examples Using Sparse Coarse Coding. *Advances in Neural Information Processing Systems*, pages 1038–1044, 1996.

[27] R. S. Sutton. *Temporal credit assignment in reinforcement learning*. PhD thesis, Department of Computer Science, University of Massachusetts, Amherst, 1984.

[28] R. S. Sutton and A. G. Barto. *Reinforcement Learning: An Introduction*. MIT Press, 1998.

[29] M. Tan. Multi-Agent Reinforcement Learning: Independent vs. Cooperative Agents. In *Proceedings of the Tenth International Conference on Machine Learning*, volume 337, 1993.

[30] K. Tumer and N. Khani. Learning from actions not taken in multiagent systems. *Advances in Complex Systems (ACS)*, 12(04):455–473, 2009.

[31] X. Wang and T. Sandholm. Reinforcement learning to play an optimal Nash equilibrium in team Markov games. *Advances in neural information processing systems*, pages 1603–1610, 2003.

[32] M. Wellman and J. Hu. Conjectural equilibrium in multiagent learning. *Machine Learning*, 33(2):179–200, 1998.

[33] E. Wiewiora. Potential-based shaping and Q-value initialization are equivalent. *Journal of Artificial Intelligence Research*, 19(1):205–208, 2003.

[34] D. Wolpert and K. Tumer. An introduction to collective intelligence. Technical Report cs.LG/9908014, NASA Ames Research Center, 1999.

[35] M. Wooldridge. *An Introduction to MultiAgent Systems*. John Wiley and Sons, 2002.