Spiking Neural Networks with Different Reinforcement Learning (RL) Schemes in a Multiagent Setting

Chris Christodoulou and Aristodemos Cleanthous

Department of Computer Science, University of Cyprus
75 Kallipoleos Avenue, P.O. Box 20537, 1678 Nicosia, Cyprus

Abstract

This paper investigates the effectiveness of spiking agents when trained with reinforcement learning (RL) in a challenging multiagent task. In particular, it explores learning through reward-modulated spike-timing dependent plasticity (STDP) and compares it to reinforcement of stochastic synaptic transmission in the general-sum game of the Iterated Prisoner's Dilemma (IPD). More specifically, a computational model is developed where we implement two spiking neural networks as two "selfish" agents learning simultaneously but independently, competing in the IPD game. The purpose of our system (or collective) is to maximise its accumulated reward in the presence of reward-driven competing agents within the collective. This can only be achieved when the agents engage in a behaviour of mutual cooperation during the IPD. Previously, we successfully applied reinforcement of stochastic synaptic transmission to the IPD game. The current study utilises reward-modulated STDP with eligibility trace and results show that the system managed to exhibit the desired behaviour by establishing mutual cooperation between the agents. It is noted that the cooperative outcome was attained after a relatively short learning period which enhanced the accumulation of reward by the system. As in our previous implementation, the successful application of the learning algorithm to the IPD becomes possible only after we extended it with additional global reinforcement signals in order to enhance competition at the neuronal level. Moreover it is also shown that learning is enhanced (as indicated by an increased IPD cooperative outcome) through: (i) strong memory for each agent (regulated by a high eligibility trace time constant) and (ii) firing irregularity produced by equipping the agents' LIF neurons with a partial somatic reset mechanism.

Key Words: spiking neural networks, multiagent reinforcement learning, reward-modulated spike timing-dependent plasticity

Introduction

Multiagent Reinforcement Learning (MARL) utilizes the most natural way of learning where the agents learn through the interaction with each other and their exploitable environment. It has been an active and intense research area over the last years where numerous successful learning algorithms have been designed; some examples include minimax-Q (11), Nash-Q (8) and FoF-Q (Friend-or-Foe Q) (12). Spiking neural networks (NNs) are computational models inspired by the structure and functionality of the neuronal system. They comprise of a collection of interconnected computational units which are modeled according to the biological neurons and communicate with each other through signals similar to neurons' action potentials. These computational systems can be employed amongst others for learning purposes. Although it naturally follows that MARL can be extended to spiking NNs as well, it is also natural to question why someone should consider using spiking neural networks with MARL in the first place. The

Corresponding author: Dr. Chris Christodoulou, Department of Computer Science, University of Cyprus 75 Kallipoleos Avenue, P. O. Box 20537, 1678 Nicosia, Cyprus. Tel: +357 22892752, Fax: +357 22892701, E-mail: cchrist@cs.ucy.ac.cy
Received: November 13, 2009; Revised: February 20, 2010; Accepted: February 23, 2010.

motivation for doing so is directly linked to the end-task of the application, which in our case was to model the high-level behaviour of self-control (3). However, this is outside the scope of this article and we will thus concentrate on the application of MARL on spiking neural networks for the purposes of a challenging game theoretical interaction. More precisely, the article presents two learning spiking neural agents that compete in the Iterated Prisoner's Dilemma (IPD).

In contrast to the case of traditional neural networks, it is only recently that reinforcement learning (RL) (19) has been successfully applied to spiking neural networks. These schemes achieve learning by utilising various biological properties of neurons whether this is neurotransmitter release (17), spike timing (6) or firing irregularity (20). The degree of experimental justification varies and it needs to be further assessed; nevertheless all these methods are biologically plausible and constitute the basis of successful RL application on biologically realistic neural models. A popular implementation of RL on spiking neural networks is achieved by modulating spike-timing-dependent synaptic plasticity (STDP) with a reward signal (5, 6, 8, 10). Other examples include Seung's reinforcement of stochastic synaptic transmission (17) as well as reinforcement of irregular spiking (20), where the learning rules perform stochastic gradient ascent on the expected reward by correlating the neurotransmitter release probability and the fluctuations in irregular spiking respectively with a reward signal. Moreover in another study, a spiking neural network implements an actor-critic Temporal Difference (TD) learning agent (15). These algorithms were shown to be able solve simple tasks like the XOR problem (6, 17). In addition, reward-modulated STDP could learn arbitrary spike patterns (5) or precise spike patterns (10) as well as temporal pattern discrimination (10) and could be used in simple credit assignment tasks (8).

Previously, we successfully applied reinforcement of stochastic synaptic transmission to a much more demanding learning task, the IPD game (3). The current study chooses to utilise Florian's algorithm which achieves RL through reward-modulated STDP with eligibility trace (6). Both algorithms (which are biologically plausible and have been derived analytically) are tested and compared with respect to their performance in the IPD. Moreover, in Florian's algorithm (6), the effect of the eligibility trace time constant, which determines the depth of memory is examined, as well as the effect of high firing variability produced by equipping the agents' leaky integrate-and-fire (LIF) neurons with a partial somatic reset mechanism (2, 4). To the best of our knowledge, this is the first time Florian's algorithm (6) is employed in a challenging multiagent task.

The IPD is a general-sum game where the payoffs' structure is such that the agents are required to exploit each other as in a fully competitive game but in a way that benefits all agents, as it would have been in a team game. The contradictory nature of these games makes their study in multiagent systems quite challenging as the agents are required to exhibit a non-trivial 'intelligent' behaviour.

In its standard one-shot version, the Prisoner's Dilemma (PD) (16) is a game summarised by the payoff matrix of Table 1. There are two competing players, Row and Column. Each player has the choice of either to "Cooperate" (C) or "Defect" (D). The "dilemma" faced by the rational players in any valid payoff structure is that, whatever the other does, each one of them is better off by defecting than cooperating. However, the outcome obtained when both defect is worse for each one of them than the outcome they would have obtained if both had cooperated. In game theoretical terms, DD is the only Nash equilibrium outcome (13), whereas only the cooperative (CC) outcome satisfies Pareto optimality (14).

The IPD is the game where the one-shot PD is played iteratively for a number of rounds. Apart from the payoff structure, the game specifies that one round of the game consists of the two players (agents) choosing their action simultaneously and independently and then informed about the outcome. It also prerequisites that the two players are rational in the sense that each player wants to maximise his or her own payoff. For the purposes of the current work we model the infinitely iterated version of the game where the same game is repeated for an unspecified amount of rounds. In the infinitely repeated version, the two agents do not have any valid reason to believe that the next round of the game will be the final one. The analysis of the infinitely repeated version of the game is much more complex than the one-shot version in terms of equilibria and successful strategies. It constitutes CC as a valid possible Nash equilibrium of the game and in addition, the CC outcome is the best possible long-term outcome both for the system in total as well as for the agents individually. The latter is true because the possible outcome where one agent always "Defects" and the other "Cooperates" can never be sustained. An extra rule (2R>T+S) (see Table 1) guarantees that the players are not collectively better off by having each player alternate between C and D, thus keeping the CC outcome Pareto optimal.

The remainder of the paper is organised as follows. The following section describes the methodology for implementing the IPD with spiking agents that learn through reward-modulated STDP. The results are presented and analysed next, while the last section discusses the results and gives the conclusions.

Table 1. Payoff matrix of the Prisoner's Dilemma, showing the specific values that are used in our simulations. For each pair of choices, the payoffs are displayed in the respective cell of the payoff matrix where payoff for the Row player is shown first. R is the "reward" for mutual cooperation. P is the "punishment" for mutual defection. T is the "temptation" for unilateral defection and S is the "sucker's" payoff for unilateral cooperation. The payoffs in any valid payoff structure should be ordered such that T>R>P>S.

	Cooperate (C)	Defect (D)
Cooperate (C)	R(=4), R(=4)	S(=-3), T(=5)
Defect (D)	T(=5), S(=-3)	P(=-2), P(=-2)

Materials and Methods

The IPD is simulated through an iterative procedure, which starts with a decision by the artificial agents, continues by feeding this information to the agents, during which learning takes place, and ends by a new decision. The agents take their first decision randomly. Each agent is implemented by a spiking neural network. The system's architecture is depicted in Fig. 1. The networks receive a common input of 60 Poisson spike trains grouped in four neural populations. Each network has a hidden layer of 60 LIF neurons and an output layer of 2 LIF neurons. The structure of the input and hidden layer as well as the equation and values of the parameters used for modelling the LIF neurons in the current implementation are the same as in (6). In addition, the output layer has only 2 LIF neurons because there are two actions that the networks choose from. Unless otherwise specified, the eligibility trace time constants are set to 25 ms, a value that is within the experimentally identified bounds.

The networks learn simultaneously but separately where each network seeks to maximise its own accumulated reward. Learning is implemented through reward-modulated STDP with eligibility trace (6) where the modulation of standard antisymmetric STDP with a reward signal leads to RL. The synaptic efficacies exhibit Hebbian STDP when the network is rewarded and anti-Hebbian when punished, allowing the network to associate an output to a given input only when accompanied by a positive reward and disassociate one when accompanied by a punishment, permitting thus the exploration of better strategies. Moreover it involves a biological plausible variable, the eligibility trace that serves as a decaying memory of the relation between recent pre- and postsynaptic spike pairs. On a previous work (3) we implemented learning through reinforcement of stochastic transmission (17) and details can be found therein. For that implementation, the equation and values of the parameters used for modelling the LIF neurons are the same as in (17) (including the eligibility trace time constants which are equal to 20 ms), apart from the value of the mean weight of the conductance used for the excitatory synapses which is set to 14nS. Both algorithms are derived as an application of the online partially observable Markov decision process (OLPOMDP) reinforcement learning (1) algorithm and also keep a record of the agents' recent actions through the eligibility trace. In reward-modulated STDP the agent is regarded to be the neuron that acts by spiking and the parameter that is optimised is its synaptic connection strengths. On the other hand, in reinforcement of stochastic transmission the synaptic connection strengths are constant, the agent is regarded to be the synapse itself that acts by releasing a neurotransmitter vesicle and the parameter that is optimised is one that regulates the release of the vesicle. Results from both algorithms are presented here as part of a more comprehensive study on MARL and spiking neural networks.

During each learning round, the networks receive a common input of 60 Poisson spike trains grouped in four neural populations which encode the decisions the two networks had during the previous round of the game. In particular, the decision of each network is encoded in the input, by the firing rate of two groups of Poisson spike trains. The first group will fire at 40 Hz if the network cooperated and at 0 Hz otherwise.

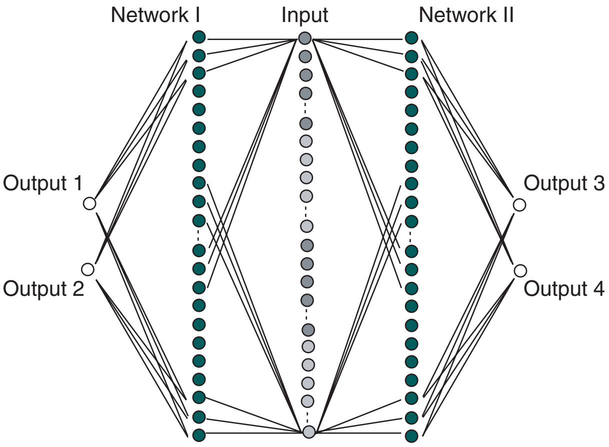

Fig. 1. Two spiking neural networks with multilayer-type architecture compete in the IPD. Each network has two layers of synapses that make full feedforward connections between three layers of neurons: the 60 shared input neurons, 60 leaky integrate-and-fire (LIF) hidden neurons and two LIF output neurons, randomly chosen to be either excitatory or inhibitory. The two networks simulate the corresponding two players of the game.

The second group will fire at 40 Hz if the network defected and at 0 Hz otherwise. Consequently, the total input to the networks during each round is represented by four groups of Poisson neurons, two groups for each network, where each group fires at 40 Hz or 0 Hz accordingly. For any given round there are always two groups of 40 Hz Poisson spike trains, preserving thus a balance at the firing rates of the output neurons at the beginning of learning. Therefore, any significant difference in the firing rate of the output neurons at any time should be induced only by learning and not due to differences in the firing rates of the driving input. It has to be noted that the frequency of 40 Hz is increased to100 Hz when the LIF neurons of the network are equipped with the partial somatic reset mechanism. A learning round lasts as long as the input is presented which is 500 ms.

At the end of each learning round the networks decide whether to cooperate or defect for the next round of the game. Decisions are carried out according to the value that each network assigns to the two actions, and these values are reflected by the firing rates of the output neurons at the end of each learning round. The value of cooperation for network *I* and *II* is taken to be proportional to the firing rate of output neurons *1* and *3* respectively. Similarly, the value of defection for network *I* and *II* is taken to be proportional to the firing rate of output neurons *2* and *4* respectively. At the end of each learning round the firing rates of the competing output neurons are compared, for each network separately, and the decisions are drawn.

When the two networks decide their play for the next round of the IPD, they each receive a distinct payoff given their actions and according to the payoff matrix of the game (Table 1). The payoff each network receives as a result of their combined actions at the previous round of the game is also the global reinforcement signal that will train the networks during the next learning round and thus guide the networks to their next decisions. For example, if the outcome of the previous round was a *CD* then according to the payoff matrix, network *I* should receive a payoff of -3 for cooperating and network *II* a payoff of +5 for defecting. During the next learning round network *I* receives a penalty of -3 and network *II* a reward of +5. The reinforcement signals are administered to the networks throughout the learning round as prescribed by the learning algorithm. Each network was reinforced for every spike of their output neuron that was “responsible” for the decision at the last round and therefore for the payoff received. Hence in the *CD* case, network *I* would receive a penalty of -3 for every spike of output neuron *1* (remember that the firing rate of output neuron *1* reflects the value that network *I* has for the action of cooperation) and network *II* would receive a reward of +5 for every spike of output neuron *4* (remember that the firing rate of output neuron *4* reflects the value that network *II* has for the action of defection). The networks therefore learn through global reinforcement signals which strengthen the value of an action that elicited reward and weaken the value of an action that resulted to a penalty.

In order to introduce competition between output neurons during a learning round, additional global reinforcement signals are administered to the networks for every spike of the output neurons that were not “responsible” for the decision at the last round. For example in the CD case, an additional reward of +1.15 is provided to network *I* for every spike of output neuron *2* and an additional penalty of -1.15 is provided to network *II* for every spike of output neuron *3*. The value of the action that was not chosen by each network is therefore also updated, by an opposite in sign reinforcement signal. The value of 1.15 is chosen to be small enough such that firstly any changes to the values of the players’ actions are primarily induced by the reinforcement signals provided according to the payoff matrix of the game and secondly, such that this complementary reinforcement signal would not cause a violation of the payoff rules that should govern the IPD.

Overall during a learning round, each network receives global, opposite in sign reinforcements for spikes of both of its output neurons. One of the two signals is due to the payoff matrix of the game and its purpose is to “encourage” or “discourage” the action that elicited reward or penalty and the other signal is complementary and is purpose is to “encourage” or “discourage” the action that could have elicited reward or penalty if had been chosen in the previous round of the game.

Results

The IPD is simulated given the system configuration described in the previous section. Each game consists of 50 rounds during which the two networks seek to maximise their individual accumulated payoff by cooperating or defecting at every round of the game. Two distinct sets of simulations were performed one for each learning scheme. The simulations aim to investigate the capability of the spiking NNs to cooperate in the IPD as well as to compare the efficiency of the two learning algorithms in the respective task. Fig. 2 shows that the implementation of the game was successful with both algorithms performing really well when the additional reinforcement signal was administered. The cooperative outcome was attained after a relatively short learning period which enhanced the accumulation of reward

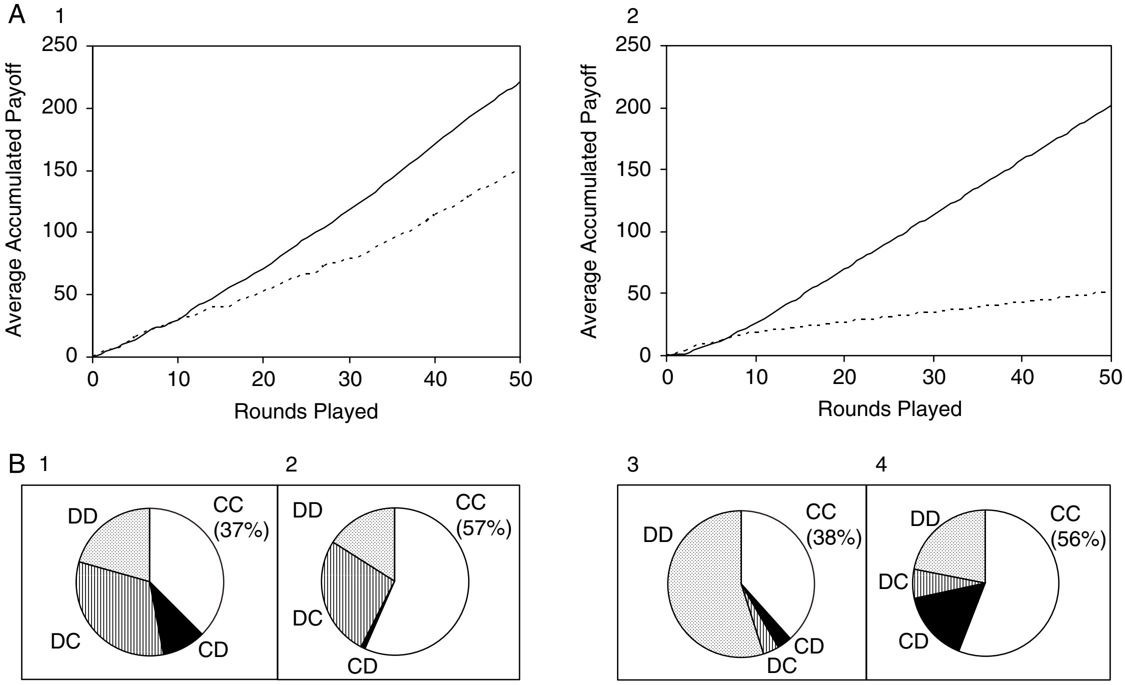

Fig. 2. **A.** The average accumulated payoff, gained by both networks as a group, during 50 rounds of the IPD. Lines represent different simulations with different global reinforcement administration routines. The solid lines represent the simulation with extra reinforcement and the dotted lines without extra reinforcement. **1.** System performance with reinforcement of stochastic synaptic transmission. **2.** System performance with reward-modulated STDP with eligibility trace. **B.** The outcomes of the IPD, given the networks' choices over the 50 rounds of the game **1.** Reinforcement of stochastic synaptic transmission without extra reinforcement. **2.** Reinforcement of stochastic synaptic transmission with extra reinforcement. **3.** Reward-modulated STDP with eligibility trace without extra reinforcement. **4.** Reward-modulated STDP with eligibility trace with extra reinforcement.

by the system. This reveals that after a certain point the networks successfully learned to resist the temptation payoff provided by defection in order to maximise their long-term reward through cooperation, enabling thus reward maximisation by the system as well. The system with reinforcement of stochastic synaptic transmission gained an average accumulated payoff of 221 where the agents cooperated 57% of the time and a payoff of 202 with 56% cooperation when trained with reward-modulated STDP. The performance deteriorated significantly when no additional reinforcement signals were administered to the networks, achieving cooperation levels of 37% and 36% respectively. Moreover, in the case of reward-modulated STDP, the total accumulated payoff decreased around 75% mainly due to high levels of mutual defection. In general, both algorithms performed the same with respect to the cooperative outcome with and without additional reinforcement signals. The important outcome of this set of simulations is that the algorithms managed to establish mutual cooperative behaviour between the agents when incorporated additional reinforcement signals.

The next set of experiments involves the more realistic learning algorithm (6) and was carried out in order to determine the effect of eligibility trace time constant (τ_z) on the ability of the networks to cooperate. The eligibility trace is a decaying memory of the relation between recent pre- and postsynaptic spike pairs and its time constant regulates the rate of this decay. Fig. 3 displays how reward-modulated STDP with additional reinforcement performed for different values of τ_z. It is shown that when both networks were configured with a weak memory, learning was totally destroyed and as a result the system received a negative accumulated payoff. The performance of the system was better when one network had strong memory and the other had weak memory, but again it was much poorer than when both networks had a strong memory. The superiority of the system with the strong memory configuration was evident. It is noted that the CC outcome not only persisted during the final rounds of the simulations, but it also did not change after a point due to the system's dynamics that were evolved by that point in time in such a way to produce CC consistently.

The final set of experiments investigates the effect of an increased firing irregularity (produced by equipping the agents' LIF neurons with a partial somatic reset mechanism) on the IPD's cooperative

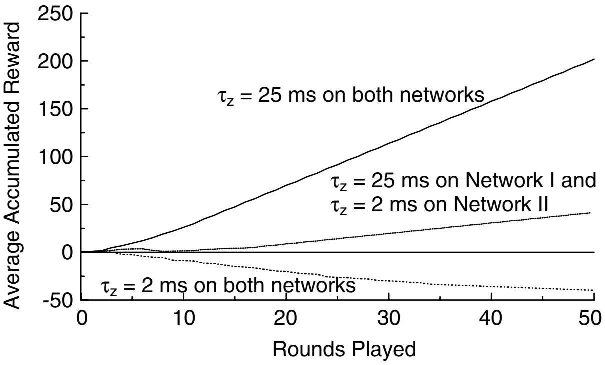

Fig. 3. The effect of eligibility trace time constant (τ_z) on the system with reward-modulated STDP. The average accumulated reward, gained by both networks as a group, during 50 rounds of the IPD. Lines represent different simulations with different eligibility trace time constants (τ_z). For all simulations the learning rate was set to 0.7.

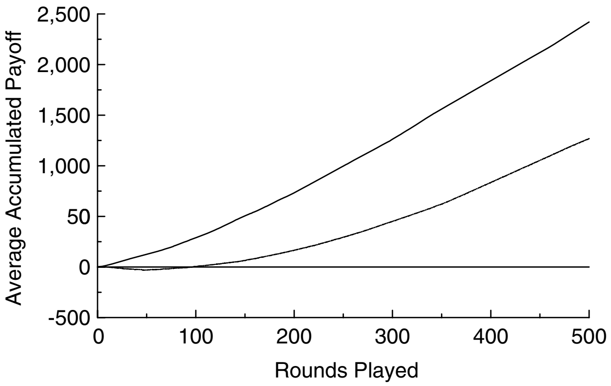

Fig. 4. Average Accumulated Reward with the LIF neurons of both networks having either partial somatic reset at 91% of the membrane potential threshold (solid line) or total reset (dotted line). Both networks learn with reward-modulated STDP with eligibility trace (6) with extra reward. The results recorded are the averages over 10 times of playing the IPD of 500 rounds each. For both networks the eligibility trace time constant is set to 25 ms and the learning rate to 0.00001; all other parameter values are as in (6).

outcome when trained with the reward-modulated STDP with eligibility trace (6). As we have previously shown (2, 4), the LIF neuron with the partial somatic reset mechanism (which corresponds to weak repolarisation of the membrane potential, see (9) for older references to this approach) is a very good candidate for producing spike trains consistent with the Poissonian output at high rates observed in the analysis of experimental recordings of cortical neuron spike trains by Softky and Koch (18).

Two simulations were performed, one with the spiking agents having LIF neurons with total reset and one with partial reset, with the partial reset level set to 91% of the membrane potential theroshold; this level of the reset parameter was chosen as it was found to produce the Poissonian firing postulated for the experimental spike trains, giving in other words high firing irregularity at high rates (2, 4). The results of both simulations are shown in Fig. 4. In this set of experiments, each game consited of 500 rounds averaged over 10 games. The difference in the system's performance is evident. Certainly with both configurations the system learns to cooperate, but when each of the competing networks of the system comprises of LIF neurons equipped with the partial somatic reset mechanism, the accumulated payoff is much higher than when there is total reset after each firing spike; this results from the difference in the cooperative outcome. With the partial reset the two networks learned quickly to reach very strong cooperation in order to maximise their long-term reward and achieved the CC outcome 64% of the time on average. On the contrary, with total reset, learning is not as strong, which is evident by the fact that the system exhibited much less cooperation (36% of the time on average).

Discussion

The current study investigated the application of RL on spiking neural networks in a demanding multiagent setting. Results showed that both investigated learning algorithms achieved to exhibit 'sophisticated intelligence' in a non-trivial task. The spiking agents showed a capacity for playing the game along the lines of game theory in a way that resembles the behaviour of real players. During most simulations, the networks managed to adapt to the challenges of the game and make decisions according to the other player's decisions in order to maximise their accumulated payoff. Most importantly, they "displayed intelligence" because when the game flow allowed for the Pareto optimum solution to be reached they "took advantage of the possibility" and settled to the solution by choosing cooperation for the rest of the game. The administration of additional global reinforcement signals, which increased competition at the neuronal and synaptic level, proved to be crucial for the high performance of the algorithms. More specifically it was essential, so as to avoid a positive feedback effect which would have increased the synaptic strength without bounds, leading to saturation of the synaptic connection and thus preventing further learning from taking place (like the limitation of classical Hebbian learning). It is noted that reward-modulated STDP performed better than reinforcement of stochastic synaptic transmission in establishing mutual cooperation between the agents of the game. Moreover, successful implementation of reward-modulated STDP required high values of eligibility trace time constants for both networks. It follows that the extent to which the reinforcement applies in changes happened before, determines the success of the learning algorithms. Results showed that reinforcement should apply to changes over a longer period of time, given that agents with a "stronger memory" configuration achieved the best cooperative result, indicating the importance of memory depth in MARL.

The increased firing irregularity at high rates, which results from the introduction of the partial somatic reset mechanism at every LIF neuron of the networks of the multiagent system, enhances the system's learning capability given the resulting accumulation of higher cooperative reward. More specifically, this firing irregularity at high rates enhances the reward-modulated STDP with eligibility trace. This could be a result of a possible increased correlation between pre- and postsynaptic spike pairs due to the high firing variability in relation with the high input frequency (100 Hz). In general, the use of LIF neurons with the partial somatic reset mechanism is very important as it models more precisely the high firing variability of cortical neurons at high firing rates (18) and as we have seen here it enhances learning as well. They could also certainly replace the artificially created firing spike trains with Poisson statistics required and used by Xie and Seung (20) for the validity of their learning rule.

Acknowledgments

We gratefully acknowledge the support of the Cyprus Research Promotion Foundation as well as the European Union Structural Funds for grant PENEK/ENISX/0308/82.

References

1. Baxter, J., Bartlett, P.L. and Weaver, L. Experiments with infinite-horizon, policy-gradient estimation. *J. Artif. Intell. Res.* 15: 351-381, 2001.
2. Bugmann, G., Christodoulou, C. and Taylor, J.G. Role of temporal integration and fluctuation detection in the highly irregular firing of a leaky integrator neuron with partial reset. *Neural Comput.* 9: 985-1000, 1997.
3. Christodoulou, C., Banfield, G. and Cleanthous, A. Self-control with spiking and non-spiking neural networks playing games. *J. Physiol.-Paris*, 104: 108-117, 2010.
4. Christodoulou, C. and Bugmann, G. Coefficient of Variation (CV) *vs*. Mean Interspike Interval (ISI) curves: what do they tell us about the brain? *Neurocomputing* 38-40: 1141-1149, 2001.
5. Faries, M.A. and Fairhall, A.L. Reinforcement learning with modulated spike timing-dependent synaptic plasticity. *J. Neurophysiol.* 98: 3648-3665, 2007.
6. Florian, R.V. Reinforcement learning through modulation of spike-timing-dependent plasticity. *Neural Comput.* 19: 1468-1502, 2007.
7. Hu, J. and Wellman, M.P. Nash Q-learning for general-sum stochastic games. *J. Machine Learning Res.* 4: 1039-1069, 2003.
8. Izhikevich, E.M. Solving the distal reward problem through linkage of STDP and dopamine signalling. *Cereb. Cortex* 17: 2443-2452, 2007.
9. Lánský, P. and Musila, M. Variable initial depolarization in Stein's neuronal model with synaptic reversal potentials. *Biol. Cybern.* 64: 285-291, 1991.
10. Legenstein, R., Pecevski, D. and Maass, W. A learning theory for reward-modulated spike-timing-dependent plasticity with application to biofeedback. *PLoS Comput. Biol.* 4: e1000180, 2008.
11. Littman, M.L. Markov games as a framework for multi-agent reinforcement learning.: In: Proc of the 11 Int Conf on Machine Learning (ICML), edited by Cohen, W. and Hirsh, H. San Francisco, PA: M. Kaufmann, pp. 157-163, 1994.
12. Littman, M.L. Friend-or-Foe Q-learning in general-sum games. In: Proc of the 18th Int Conf on Machine Learning (ICML), edited by Brodley, C., Danyluk, A. San Francisco, PA: M. Kaufmann, pp. 322-328, 2001.
13. Nash, J. Equilibrium points in N-person games. *Proc. Natl. Acad. Sin. U.S.A.* 36: 48-49, 1950.
14. Pareto, V. Manuale di economia politica. Milan: Societa Editrice, 1906.
15. Potjans, W., Morrison, A. and Diesmann, M. A spiking neural network model of an actor-critic learning agent. *Neural Comput.* 21: 301-339, 2009.
16. Rappoport, A. and Chammah, A.M. Prisoner's dilemma: a study in conflict and cooperation. Ann Arbor, MI, USA: Univ. of Michigan Press, 1965.
17. Seung, H.S. Learning in spiking neural networks by reinforcement of stochastic synaptic transmission. *Neuron* 40: 1063-1073, 2003.
18. Softky, W.R. and Koch, C. The highly irregular firing of cortical cells is inconsistent with temporal integration of random EPSPs. *J. Neurosci.* 13: 334-350, 1993.
19. Sutton, R.S. and Barto, A.G. Reinforcement Learning: An Introduction. Cambridge, MA, USA: MIT Press, 1998.
20. Xie, X. and Seung, H.S. Learning in neural networks by reinforcement of irregular spiking. *Phys. Rev. E.* 69: 41909, 2004.