Learning Aquatic Locomotion with Animats

Dennis G. Wilson, Jean Disset, Sylvain Cussat-Blanc, Yves Duthen, Hervé Luga

Learning Aquatic Locomotion with Animats

Dennis G Wilson, Jean Disset, Sylvain Cussat-Blanc, Yves Duthen, Hervé Luga

IRIT - Université de Toulouse - CNRS - UMR5505, Toulouse, France 31062
dennis.wilson@irit.fr

Abstract

One of the challenges of researching spiking neural networks (SNN) is translation from temporal spiking behavior to classic controller output. While many encoding schemes exist to facilitate this translation, there are few benchmarks for neural networks that inherently utilize a temporal controller. In this work, we consider the common reinforcement problem of animat locomotion in an environment suited for evaluating SNNs. Using this problem, we explore novel methods of reward distribution as they impacts learning. Hebbian learning, in the form of spike time dependent plasticity (STDP), is modulated by a dopamine signal and affected by reward-induced neural activity. Different reward strategies are parameterized and the Covariance Matrix Adaptation Evolution Strategy (CMA-ES) is used to find the best strategies for fixed animat morphologies. The contribution of this work is two-fold: to cast the problem of animat locomotion in a form directly applicable to simple temporal controllers, and to demonstrate novel methods for reward modulated Hebbian learning.

1 Introduction

As more of the biologic mechanisms of learning are explored, computational models are created for the advancement of artificial learning and for a deeper understanding into each mechanism's role in learning, both artificial and biologic. One such example is synaptic plasticity in neural networks, often implemented in spiking neural networks (SNN) as spike time dependent plasticity (STDP). STDP has been shown to be a powerful Hebbian learning rule, competing with other artificial intelligence (AI) methods on common benchmarks like handwriting and image recognition (Kheradpisheh et al., 2016). The use and analysis of computational STDP models has benefited both computer scientists and biologists alike, allowing for a mathematical understanding of some of the emergent behavior in the complex system of a neural network (Nessler et al., 2013).

However, there is still much to be understood regarding learning, both in biologic and artificial neural networks. For example, the physical location of synapses on a dendrite and the location of the corresponding neuron in the brain may both play a part in Hebbian behavior in biologic networks (Hosp et al., 2011), but most artificial neural networks (ANN) have no representation of physical space.

Furthermore, while SNNs have clear benefits in the pursuit of understanding learning, given their biologic basis, established unsupervised and supervised learning rules, and wealth of temporal information, their evaluation can be difficult when compared to other controllers that don't require time simulation. Inputs to an SNN are often cast as random distributions with parameters set by the problem and a variety of encoding schemes exist to translate the temporal output of an SNN to a scalar output necessary for classification or many control problems Grüning and Bohte (2014). Simple problems, such as learning a specific spiking sequence, are often used in place of more rigorous benchmarks; one such problem is used in this work to both demonstrate this analysis and highlight its shortcomings.

In this work, the mechanisms of reward as they impact learning are examined in an animat locomotion problem. While locomotion is a common problem, we create an environment suited to any controller that can make binary decisions in a temporal pattern, here being the output firing events of an SNN. This problem is used to evaluate existing reward modulated STDP methods and two novel model additions: the simulation of the dopaminergic signal as a chemical that propagates through time and disappates over space, and the direct activation of neurons based on extracellular dopamine. These methods are compared and the results evaluated to address the main question of this work: what are the benefits of different reward mechanisms in an SNN?

This paper is organized as follows. The foundations of STDP and similar works are described in § 2 and are expanded upon in § 3. The details of the reward methods examined in this work follows in § 4. A simple experiment is described in § 5, and the proposed animat

locomotion problem is presented in § 6, with analysis of the results of this experiment in § 7. General discussion of the results, including their main conclusions and possible improvements for future work, are detailed in § 8.

2 Related work

The SNN model has numerous forms across a variety of simulation tools (Brette et al., 2007). In its simplest form, an SNN operates by emulating the spiking activity of a neural network, first feeding an input signal to a group of input neurons which then send signals to downstream neurons upon activation. This neural activation, or spiking, corresponds to thresholds in the accumulation of input signals over time, often represented as limits on the neural membrane potential. Various encoding schemes exist to interpret output firing patterns as scalar information, such as counting the number of spikes in a given window or using the first spike from a set of neurons (Grüning and Bohte, 2014). While some problems exist that directly utilize the temporal nature of an SNN, such as the temporal information processing in (Kasabov et al., 2013), there is a lack of fitting problems from other domains, particularly reinforcement learning.

With the use of supervised learning methods, SNNs have displayed impressive results. ReSuMe (Ponulak, 2005) uses external training neurons to modify the weights of an SNN in order to produce a desired spiking output sequence. Deep learning methods have also been applied to SNNs with success: in (Hunsberger and Eliasmith, 2015), a deep neural network is trained and then converted into an SNN that performs competitively on the MNIST and CIFAR-10 recognition benchmarks.

STDP has also shown impressive capabilities in unsupervised learning tasks, such as clustering. In (Diehl and Cook, 2015), STDP is used to train a network capable of differentiating the handwritten digits of MNIST with high accuracy. Similarly, (Kheradpisheh et al., 2016) uses STDP on a number of recognition tasks, including MNIST, and performs competitively against other standard methods.

Between the supervised methods with detailed error information, commonly in the form of a desired output sequence, and the unsupervised clustering of STDP, which trains only on input information, lie semi-supervised methods. In (Kasabov et al., 2013), a combination of a variant of STDP, spike driven synaptic plasticity (SDSP), and supervised learning are used on an EEG pattern recognition task. In semi-supervised implementations of SDSP, a learning signal is applied along with the problem input signal; this change in network activity then influences SDSP and therefore the weight modification.

STDP is also directly modified by output reward signals, such as in Farries and Fairhall (2007), where a multiplicative factor based on reward is introduced in the weight update rule. Various methods for modulating STDP are reviewed in (Frémaux and Gerstner, 2016). Inspired by biologic dopamine signaling, (Izhikevich, 2007) expands the concept of reward modulation by using a dopamine concentration STDP modulation coefficient that decays over time (DA-STDP). While this shows impressive results in attributing delayed rewards to the appropriate synapses, the implementation and evaluation of this model on a classic problem was shown to be challenging in Chorley and Seth (2008). DA-STDP is evaluated in this work, both on one of the problems from (Izhikevich, 2007) and on the proposed novel benchmark.

The proposed locomotion problem is used to evaluate methods from (Frémaux and Gerstner, 2016) and two novel method additions. Similar problems have been used to study evolved biologic controllers, such as in (Joachimczak et al., 2016), where 2D animats develop a gait by controlling the expansion and contraction of their cells. This work further tunes this type of problem for use with SNNs by fixing the contraction event to fit binary spiking output. The animat and its 3D environment are encoded using the integrated physics engine of the artificial life platform, MecaCell (Disset et al., 2016).

3 Spiking Neural Networks

In this work, the Izhikevich SNN model is used (Izhikevich, 2004), as it can exhibit a variety of natural behaviors. In this model, each neuron n has a membrane potential v_n and a membrane recovery variable u_n. v_n is increased by input I_n either from external sources or from other neurons:

$$v_i(t=0) = v_R$$
$$v_i(t+1) = 0.04v_i(t)^2 + 5v_i(t) + 140 - u_i(t) + I_i(t)$$
$$u_i(t+1) = a(bv_i(t) - u_i(t))$$

The membrane potential is increased from a resting potential v_R until reaching a threshold v_T, at which point the neuron spikes, resetting v to a membrane potential c and updating u. A signal from the spiking neuron then propagates to post-synaptic neurons, increasing their synaptic input I_j by the weight s from the spiking neuron n_i to the post-synaptic neuron n_j:

$$v_i(t+1) = c \;\; ; \;\; u_i(t+1) = u_i(t) + d$$
$$I_j(t+1) = I_j(t+1) + s_{i,j}$$

Synapses are modeled as a matrix of real valued weights; excitatory synapses are initialized to s_e and

v_T	30.0	v_R	-65.0
a	(0.02, 0.1)	b	0.2
c	-65.0	d	(8.0, 2.0)
s_e	1.0	s_i	-1.0
A_+	1.0	A_-	1.5

Table 1: Neural parameters from (Izhikevich, 2007). The two values of a and d correspond to excitatory and inhibitory neurons, respectively.

bound between $[0.0, 4.0]$ during STDP training , and inhibitory synapses are held constant at s_i.

3.1 Spike Time Dependent Plasticity

STDP modifies the synaptic weights of an SNN based on the fire timing of the synapses's respective neural endpoints. If a neuron n_i fires and then a post-synaptic neuron n_j fires shortly after, the synaptic weight $s_{i,j}$ is increased. If the firing order is reversed, $s_{i,j}$ is decreased. Using this Hebbian learning scheme, hidden neurons are tuned to features in the input layer, as captured visually in (Diehl and Cook, 2015). Many STDP schemes use a neural competition rule, such as in (Kheradpisheh et al., 2016), where the first neuron in any layer to fire is the only one trained for a given input sequence. In this work, no fixed competition rule is used; instead, it is through the distribution of reward that STDP applies variably to competing neurons. The STDP update rule from (Izhikevich, 2007) is used:

$$\Delta s_{i,j} = A_+ e^{-(t_j - t_i)} \delta(t - t_i), \quad if \quad t_j - t_i > 0$$
$$\Delta s_{i,j} = -A_- e^{-(t_j - t_i)} \delta(t - t_i), \quad if \quad t_j - t_i < 0$$

where t_i indicates the most recent spike time of neuron n_i, A_+ and A_- are STDP learning parameters, and $\delta(t)$ is the Dirac delta function. Euler integration with a 1ms time step is used for computation.

4 Reward methods

In this paper, two semi-supervised learning methods are expanded upon and parameterized for exploration. Both methods build on STDP using an artifical dopamine concentration, which is a function of a global reward signal, rw, provided to the controller. The dopamine concentration is calculated for each neuron, dependent on its position. Neurons in this work are positioned in a 3D space, with topology determined per task. A dopamine signal starts at the center of mass of the network and propagates at a speed based on a delay parameter, p_{dd}. The concentration of this signal also attenuates as it travels based on the parameter p_{dat}. Lastly, a fraction of dopamine concentration is absorbed each timestep based on p_{dab}:

$$dist = \frac{\sqrt{\sum_{d=0}^{2}(n_i[d] - com[d])^2}}{dist_{max}}$$
$$dx = p_{dd} size(h) dist$$
$$r_i = rw[\lfloor dx \rfloor] + (dx - \lfloor dx \rfloor)(rw[\lfloor dx \rfloor + 1] - rw[\lfloor dx \rfloor])$$
$$D_i(t+1) = (1.0 - p_{dab}) D_i(t) + e^{-p_{dat} * dist} r_i$$

where com is the network's center of mass, rw is a fixed-size array of the most recent reward values calculated at a fixed interval, d indicates the positional dimension, and $dist_{max}$ is the maximum radius of the network. The dopamine concentration at each neuron is therefore a scaled version of the linear interpolation of the delayed reward based on the propagation delay, p_{dd}, with accumulation over time based on p_{da}.

The first reward method proposed is the direct input of reward as an activation mechanism, termed Induced Firing STDP (IF-STDP). This method is based on dopaminergic activation in biologic brains (Pignatelli and Bonci, 2015) and has similarities to the teaching neurons of semi-supervised methods such as ReSuMe (Ponulak, 2005). IF-STDP functions by directly activating neurons based on the extracellular dopamine at their position:

$$I_j = I_j + p_{rs} D_j$$

where p_{rs} is a reward signal coefficient parameter. This is intended to induce firing based on a reward signal, which will then further strengthen the synapse between activated neurons through basic STDP.

The second reward method is the modulation of STDP using the dopamine concentration, Dopamine Modulated STDP (DM-STDP). The synaptic weight change of STDP is modified by the dopamine concentration of the involved neurons n_i and n_j, based on p_{df}, a dopamine factor parameter:

$$D_{i,j} = \frac{D_i + D_j}{2.0}$$
$$\Delta s_{i,j} = (0.01(1 - p_{df}) + p_{pd} D_{i,j}) \Delta s_{i,j}$$

By using different values of the parameters p_{rs}, p_{df}, p_{dd}, p_{dat}, and p_{dab}, different STDP modulation methods can be recreated. Classic STDP is achieved when no reward induced firing or modulation take place, hence p_{rs} and p_{df} must both be 0. R-STDP, defined in (Frémaux and Gerstner, 2016) as "gated Hebbian learning", modulates STDP based on instantaneous reward only, therefore the absorption rate parameter p_{dab} is 1.0; all dopamine is absorbed at each timestep. The signal does not travel over distance or attenuate over time, so both p_{dd} and p_{dat} are 0.0. This is the same as DA-STDP (Izhikevich, 2007), which added the novel concept of

method	p_{rs}	p_{df}	p_{dd}	p_{dat}	p_{dab}
STDP	0.0	0.0	0.0	0.0	1.0
R-STDP	0.0	**1.0**	0.0	0.0	1.0
DA-STDP	0.0	**1.0**	0.0	0.0	**0.001**
IF-STDP	**1.0**	0.0	**1.0**	**0.1**	**0.001**
DM-STDP	0.0	**1.0**	**1.0**	**0.1**	**0.001**
IFDM-STDP	**1.0**	**1.0**	**1.0**	**0.1**	**0.001**

Table 2: Parameterization of different reward modulation methods. Bold values are tunable within the method; the values for these parameters were chosen for the instrumental conditioning experiment defined in § 5

dopamine absorption over time, here reflected in p_{dab}. These parameters are given fully in Table 2.

The new methods proposed in this work, IF-STDP, DM-STDP, and their combination IFDM-STDP, use a chemical dopamine signal that propagates through the physical space of the network, attenuating as it travels and being absorbed over time. While the underlying effects of both methods are not novel, their use of such a dopamine signal is new. We therefore present both a comparison of these new methods with previous ones and an exploration of the parameter space that defines the dopamine signal and its use.

5 Instrumental conditioning

First, to display simply the functionality of each model, the instrumental conditioning experiment from (Izhikevich, 2007) was reproduced. This sort of experiment is common in SNN literature, as it focuses on specific spike timing and not on any application of the spiking output of the network.

For this experiment, a network with N_{in} input, N_h neurons, and N_{out} output neurons was used. This input, hidden, output designation does not indicate topology as is common in other ANN literature, but rather the use of the neuron. Input neurons receive a stimulus current of ϕ_s mV every ϕ_i ms.

The topology of the network is random: each neuron has a ρ_c chance of connecting with another neuron and a ρ_{in} chance of being an inhibitory neuron. Therefore, in this experiment, the network was composed of 1000 total neurons, 800 being excitatory and 200 being inhibitory, with 100 synaptic connections each. The neurons were placed randomly in a 3D space following a uniform distribution over $[-1.0, 1.0]$ in each dimension. Parameters for this experiment can be found in Table 3.

The output neurons were split into two exclusive groups of 50 neurons each, A and B. For a short period, 20ms, after each input stimulus ϕ, the number of spikes in each group was recorded as $|A|$ and $|B|$. For the first 400 stimulus intervals, a constant reward rw was provided following the stimulus if $|A| > |B|$, and when $|B| > |A|$ for the second 400 stimulus intervals. The reward was delayed by a maximum of 1s and the stimulus intervals were 10s apart.

N_{in}	50	N_h	850	N_{out}	100
ρ_c	0.1	ρ_{in}	0.2	rw	0.1
ϕ_g	1	ϕ_i	1000	ϕ_s	15

Table 3: Parameters used in the instrumental conditioning experiment. ϕ_g indicates the number of distinct input groups, which in this experiment was 1, meaning the inputs were not subdivided.

This task is difficult because the reward is delayed and is therefore challenging to correlate with the firing events that caused it. Furthermore, the goal changes after 400 intervals, requiring the weights between the input and A to decay during this second interval.

5.1 Results

In the instrumental conditioning experiment (Figure 1), DA-STDP displayed its capabilities as in (Izhikevich, 2007). While the delayed reward in this problem is difficult to properly assign, by introducing an absorption rate p_{dab}, the dopamine concentration is able to last until further firing episodes between the inputs and the output group occurred, triggering STDP. Over many cycles of stimulation and reward, the events become correlated and the synaptic weights between the inputs and the rewarded group increased.

R-STDP is not able to solve this problem due to the instantaneous gating of STDP it performs. The random delay does not allow it to correlate the reward with the proper firing events, so the weights from the inputs to both groups are increased. This is also seen in IF-STDP, where the induced firing alone is not enough to influence the weights. However, when combined with DM-STDP, some improvement is seen in IFDM-STDP; the gap between the weights widens and is reached faster. DM-STDP is a reduction in total weight change from DA-STDP, due to the dopamine attenuation p_{dat}, and the induced firing from IF-STDP appears to match DM-STDP with DA-STDP for total weight change.

While this task is challenging, the different STDP methods either fail or succeed at the task, and it is difficult to discern their quality. Furthermore, the application is only abstract; the different output groups can be considered different motor responses, but it is not clear what the output firing corresponds to or what the delayed response should represent. The random assignment of neural positions also reflects the abstract nature of this experiment; to fully explore the impact

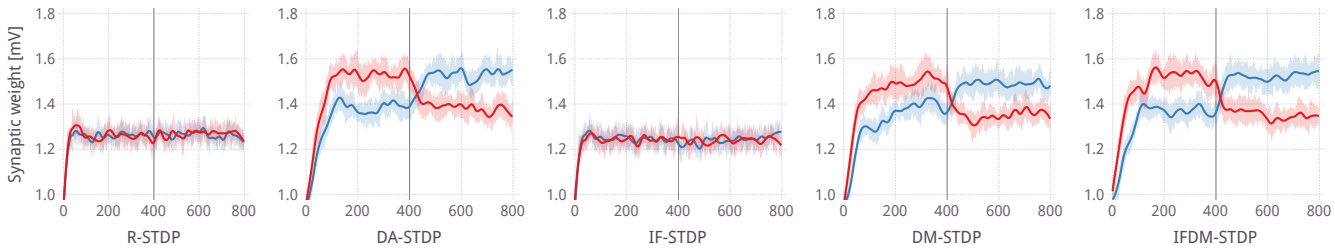

Figure 1: The synaptic weights between the input and the A (red) and B, (blue) groups in the instrumental conditioning experiment. By rewarding firing from group A during the first 400 episodes, the weights between the input and A should increase, as with group B for the second 400 episodes. Ribbons represent the standard deviation over 20 trials.

of physical parameters in an embodied network, specific topologies must be considered. To address these issues, we propose the following benchmark problem, animat locomotion, and present an exploration of the parameterized methods using evolutionary search.

6 Aquatic Locomotion Problem

The locomotion problem is classic and has many interpretations, such as that in (Joachimczak et al., 2016). Here we focused on an animat composed of linked cells which propelled itself in an aquatic environment by contracting its cells in coordinated motion. This allows for the simplistic output of synaptic firing, a binary event, to be used for control in a complex environment. Not only did the animat have to learn to coordinate firing events to create large-scale body movement, it had to do so in an advantageous way based on the fluid dynamics present and its morphology.

An SNN with specified STDP parameters was placed inside an animat with all input and hidden cells located at the animat's center of mass. Clusters of cells were controlled by their nearest output neuron, which were placed evenly throughout the morphology. Upon firing, an output neuron caused its connected cluster to contract, leading to deformation of the animat, allowing for locomotion. Input neurons were separated into ϕ_g groups and given a stimulus signal ϕ_s every ϕ_i ms, with the chosen input group rotating each stimulus.

Two static morphologies were used in this experiment to diversify the neural topologies and movement strategies. These morphologies were a four-legged octopus (quadropus) and a stingray, shown in Figure 3 and Figure 2.

Reward was initially provided to the animat based on the movement of its center of mass *com*. While this constant reward signal is desirable in many reinforcement learning settings, we found that discrete reward events were more suitable in this problem. The reward was therefore the percentage increase of animat velocity

N_{in}	440	N_h	570	N_{out}	(87, 74)
ρ_c	0.13	ρ_{in}	0.20	rw	10.0
ϕ_g	6	ϕ_i	90	ϕ_s	45
T_{cont}	20	c_{cont}	0.9	F_{fluid}	0.0005

Table 4: Parameters used in the aquatic locomotion experiment, where the two values for N_{out} correspond to the quadropus and stingray morphologies, respectively

whenever the velocity eclipsed its previous maximum, v_{max}. To continue to reward velocity increases over the life of the animat, v_{max} decayed exponentially.

$$dist(t) = \sqrt{\sum_{d=0}^{2}(com(t)[d] - com(t=0)[d])^2}$$
$$v(t) = dist(t) - dist(t-1)$$
$$v_{max}(t) = 0.99 v_{max}(t-1)$$
$$rew(t) = rw * max(0.0, (v(t) - v_{max}(t))/v_{max}(t))$$

The goal for STDP was to therefore correlate input stimulus firing with output behavior that increased velocity, similar to the instrumental conditioning experiment. Unlike that experiment, however, the mapping between output firing and reward was highly complex, as the animat had to continuously find new output firing patterns that increased its velocity.

6.1 MecaCell

We based our experiments on the Artificial Life platform MecaCell in which we created an aquatic environment. The organism was composed of several tightly packed cells linked with elastic bonds, using a mass-spring-damper system for modelling both the adhesions and the collisions. The bonds were created between neighbouring cells at the beginning of the simulation

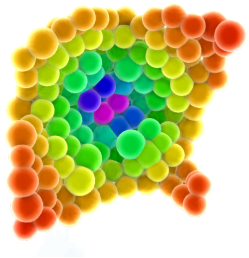

Figure 2: The stingray morphology with coloring based on the dopamine distribution

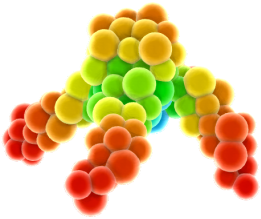

Figure 3: The quadropus morphology with coloring based on the dopamine distribution; the quadropus has a fourth arm which is obscured in this image

and were then set to be unbreakable. In order to obtain the creatures shapes, we used 3D meshes which we filled with cells.

Each cell contracted by changing its desired radius to c_{cont} times its original radius, which amounts to shortening the rest length of the collision springs and thus pulling on connected bonds. After a set duration T_{cont}, the cell reset to its original spring length. If an output neuron fired for a previously contracted cell, the cell remained contracted for another T_{cont} ms.

To encourage the use of this problem, we have made the source code available, along with videos of the best individuals of each morphology.[1] MecaCell is an open source C++ platform; this work extends MecaCell with plugins for the SNN and reward mechanism. We also used the Julia language for analysis, CMA-ES, and the instrumental conditioning experiment.

6.2 Parameter evolution

To fully explore the different STDP modulation methods, the method parameters p_{rs}, p_{df}, p_{dd}, p_{dat}, and p_{dab} were evolved using the Covariance Matrix Adaption Evolutionary Strategy (CMA-ES), a popular search algorithm for real valued numbers. The method parameters were optimized within the ranges given in Table 5, and the evolutionary fitness for maximization was the cumulative sum of the distance traveled away from the center of mass.

[1] `https://github.com/d9w/lala`

	p_{rs}	p_{df}	p_{dd}	p_{dat}	p_{dab}
min	0.0	0.0	0.0	0.0	0.0
max	2.0	1.0	1.0	10.0	1.0

Table 5: Reward parameters ranges for CMA-ES

$$dist(t) = \sqrt{\sum_{d=0}^{2}(com(t)[d] - com(t=0)[d])^2}$$

$$fitness = \sum_{t} dist(t)$$

R-STDP, DA-STDP, and IFDM-STDP were optimized independently by fixing the non-tunable parameters and optimizing the others. The population size λ for CMA-ES was chosen as a function of the parameter space size P: $\lambda = 4 + \lfloor 3log(P) \rfloor$.

CMA-ES was run for 50 evaluations and 20 independent trials were conducted for statistical testing. All parameters were optimized within $[0.0, 1.0]$ and then scaled to their respective ranges for fitness evaluation. Uniform random values were used to initialize CMA-ES and the step size for all parameters was 0.5.

7 Results

By evolving the parameters using CMA-ES, significant improvement in the total distance traveled was achieved, especially for the quadropus morphology, as seen in Figure 4. Neither R-STDP nor DA-STDP reached the distances that IFDM-STDP was able to, indicating the important of a physically situated dopamine concentration for this problem. As the network topology is directly representative of the animat morphology, with output neurons positioned throughout the animat, having a dopamine signal with delayed propagation and attenuation appears to have been very important.

To understand which parameters are responsible for the success of IFDM-STDP, the parameters of the best individuals are shown in Figure 5 as the normalized parameter values, before they are set to the parameter ranges in Table 5. The values of the parameters of single best individual are also show in Table 6. Also shown in this table are the best individuals from the evolution of R-STDP and DA-STDP.

Some parameters confer a consistent benefit. The dopamine factor p_{df} is high for all top IFDM-STDP individuals, as is the dopamine decay parameter p_{dd}. First, this that STDP utilized the dopamine concentration to modify weights. That alone is not sufficient, though, as demonstrated by R-STDP's performance.

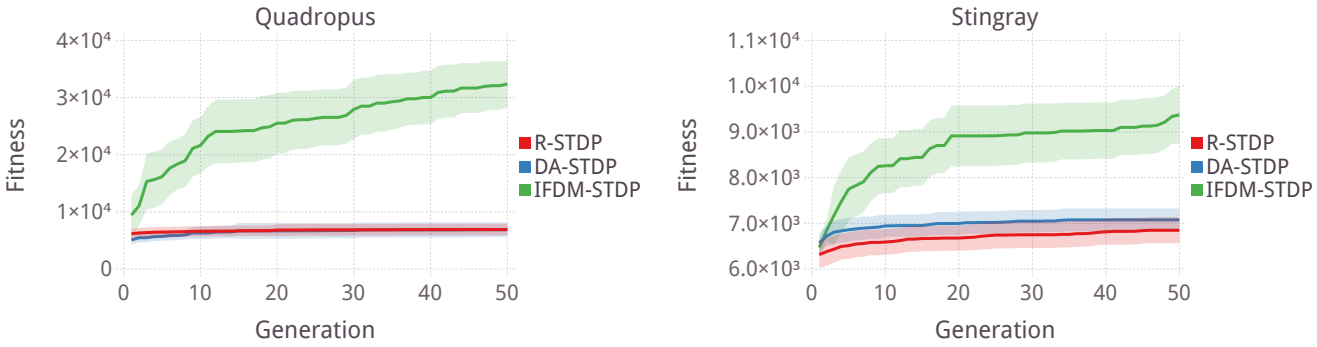

Figure 4: CMA-ES optimization of the tuneable parameters of the different STDP strategies for both morphologies. Ribbons indicate standard deviation over 20 trials.

Quadropus

method	p_{rs}	p_{df}	p_{dd}	p_{dat}	p_{dab}
R-STDP	0.0	**0.80**	0.0	0.0	1.0
DA-STDP	0.0	**0.68**	0.0	0.0	**0.95**
IFDM-STDP	**0.04**	**0.79**	**0.54**	**1.92**	**0.78**

Stingray

method	p_{rs}	p_{df}	p_{dd}	p_{dat}	p_{dab}
R-STDP	0.0	**0.69**	0.0	0.0	1.0
DA-STDP	0.0	**0.08**	0.0	0.0	**0.91**
IFDM-STDP	**1.82**	**0.83**	**0.86**	**0.80**	**0.67**

Table 6: Evolved parameters for each method on both morphologies. Bold values indicate the best evolved value, while non-bold values were held constant.

The usage of the p_{dd} parameter means that delaying the reward signal to the distal parts of the animat morphology was beneficial. As contraction events near the center of the animat often caused movement in the distal parts of the morphology, but not vice versa, this delay can be seen as a way to properly correlate reward with firing events in the distal regions and not with motion caused by central contractions.

Other parameters are not consistent between the morphologies. The reward signal factor p_{rs} was not used by most top quadropus individuals, but was by top stingray individuals. One possible explanation for this is that the quadropus is more rigid than the stingray, and excess firing can more easily have a negative effect on the movement pattern of the quadropus than the stingray. Neither morphology had a consistent strategy concerning p_{dat} either; while both best individuals had relatively low attenuation parameters, the distribution over the top individuals is wide.

8 Discussion

Given the increase in evolutionary fitness by modifying the method parameters, it is clear that some of the proposed reward mechanisms provide benefits in this problem. These benefits have been explored in the context of this work, but future work is necessary to continue to assess their impact in different settings. Specifically, these methods should be assessed in other problems in which the neural network is situated within the controlled object, giving each neuron a position in space.

Furthermore, many assumptions made in this work can be challenged. The dopamine signal for both experiments originated at the network's center of mass, but biologic dopamine signals have multiple origins and do not diffuse equally throughout the brain. Whether or not this is the product of biologic design or a feature of learning can be explored.

The learning feature of delayed reward information, here found in both p_{dab} and p_{dd}, is one that is being explored in artificial learning. The abstraction of dopamine delay can be taken from this model and used even in networks that don't have neural positioning, as long as some delay coordinate, such as layer depth, is provided. This can serve many training methods on problems with temporal reward, especially in the presence of a delay between the action and the reward.

Lastly, the locomotion problem presented in this work can be used to evaluate other methods and can be expanded upon. Preliminary trials with an evolved Gene Regulatory Network (GRN) have shown further possibilities for locomotion strategies in the same animats used in this work. In the future, we hope to use this problem with animat morphologies that develop in parallel to their controlling neural network, solving more complex tasks in the environment, such as foraging.

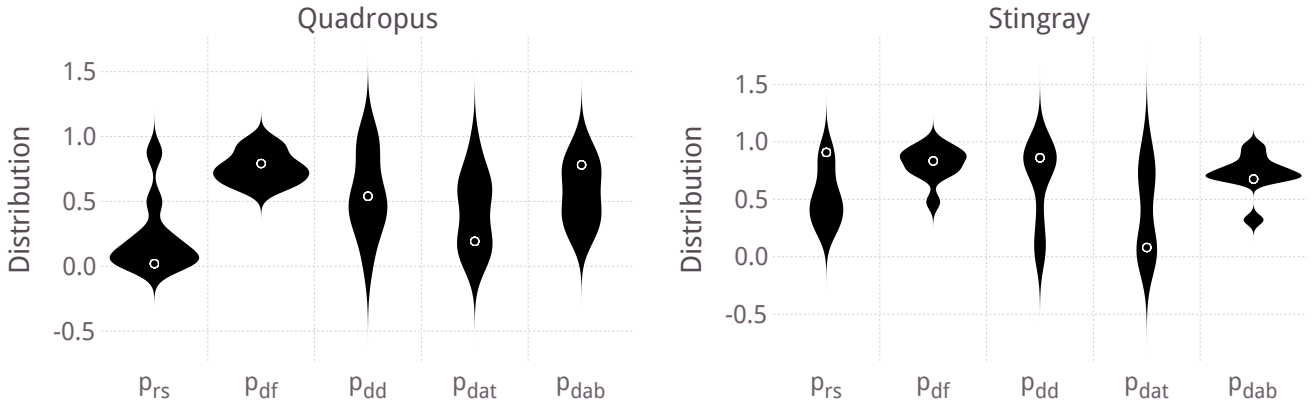

Figure 5: Distribution of the reward parameters of the 10 best individuals for both morphologies. Circles show the best single individual.

Acknowledgments

This work is supported by ANR-11-LABX-0040-CIMI, within programme ANR-11-IDEX-0002-02. This work was performed using HPC resources from CALMIP (Grant P16043).

References

Brette, R., Rudolph, M., Carnevale, T., Hines, M., Beeman, D., Bower, J. M., Diesmann, M., Morrison, A., Goodman, P. H., Harris, F. C., et al. (2007). Simulation of networks of spiking neurons: a review of tools and strategies. *Journal of computational neuroscience*, 23(3):349–398.

Chorley, P. and Seth, A. K. (2008). Closing the sensory-motor loop on dopamine signalled reinforcement learning. *Lecture Notes in Computer Science (including subseries Lecture Notes in Artificial Intelligence and Lecture Notes in Bioinformatics)*, 5040 LNAI:280–290.

Diehl, P. U. and Cook, M. (2015). Unsupervised learning of digit recognition using spike-timing-dependent plasticity. *Frontiers in computational neuroscience*, 9:99.

Disset, J., Cussat-Blanc, S., and Duthen, Y. (2016). Evolved developmental strategies of artificial multicellular organisms. *15th International Symposium on the Synthesis and Simulation of Living Systems (ALIFE XV 2016)*.

Farries, M. A. and Fairhall, A. L. (2007). Reinforcement Learning With Modulated Spike Timing-Dependent Synaptic Plasticity. *Journal of neurophysiology*, 98(6):3648–3665.

Frémaux, N. and Gerstner, W. (2016). Neuromodulated Spike-Timing-Dependent Plasticity, and Theory of Three-Factor Learning Rules. *Frontiers in Neural Circuits*, 9(January).

Grüning, A. and Bohte, S. M. (2014). Spiking Neural Networks: Principles and Challenges. *ESANN*, (April):23–25.

Hosp, J. A., Pekanovic, A., Rioult-Pedotti, M. S., and Luft, A. R. (2011). Dopaminergic projections from midbrain to primary motor cortex mediate motor skill learning. *Journal of Neuroscience*, 31(7):2481–2487.

Hunsberger, E. and Eliasmith, C. (2015). Spiking deep networks with lif neurons. *arXiv preprint arXiv:1510.08829*.

Izhikevich, E. M. (2004). Which model to use for cortical spiking neurons? *IEEE Transactions on Neural Networks*, 15(5):1063–1070.

Izhikevich, E. M. (2007). Solving the distal reward problem through linkage of STDP and dopamine signaling. *Cerebral Cortex*, 17(10):2443–2452.

Joachimczak, M., Kaur, R., Suzuki, R., and Arita, T. (2016). Spiral autowaves as minimal, distributed gait controllers for soft-bodied animats. In *Proceedings of the Artificial Life Conference 2016*, pages 140–141. MIT Press.

Kasabov, N., Dhoble, K., Nuntalid, N., and Indiveri, G. (2013). Dynamic evolving spiking neural networks for on-line spatio- and spectro-temporal pattern recognition. *Neural Networks*, 41(1995):188–201.

Kheradpisheh, S. R., Ganjtabesh, M., Thorpe, S. J., and Masquelier, T. (2016). Stdp-based spiking deep neural networks for object recognition. *arXiv preprint arXiv:1611.01421*.

Nessler, B., Pfeiffer, M., Buesing, L., and Maass, W. (2013). Bayesian Computation Emerges in Generic Cortical Microcircuits through Spike-Timing-Dependent Plasticity. *PLoS Computational Biology*, 9(4).

Pignatelli, M. and Bonci, A. (2015). Role of Dopamine Neurons in Reward and Aversion: A Synaptic Plasticity Perspective. *Neuron*, 86(5):1145–1157.

Ponulak, F. (2005). ReSuMe-new supervised learning method for Spiking Neural Networks. *Inst. Control Information Engineering, Poznan Univ.*, 22(2):467–510.