Learning Motor Policies with Time Continuous Neural Networks

Titouan Renard [1]

[1] *MT-RO, 272257*

Abstract *Time-Continuous Neural Networks provide an effective framework for the modeling of dynamical systems [8] and are natural candidates for continuous-control tasks. They are closely related to the dynamics of non-spiking neurons, which gives further justification to investigate their use in control [14]. The following document describes the work and results obtained while investigating the use of such neural networks during a semester project jointly supervised by EPFL's BIOROB and LCN labs.*

Introduction

We first consider time-continuous neural networks, and the main ideas of reinforcement learning relevant to this kind of task, then we discuss implementation details, and finally we go through the results obtained during the project.

1 Time-Continuous Neural Networks

1.1 Formulating a Neural Network model for Continuous-Time Processes

In the following discussion, we only worry about time-independent (sometimes also refered to as autonomous) systems, as those are more relevant for control, but most of those approaches generalize well to time-dependent systems as well. Time-Continuous Neural Networks model dynamical systems model the evolution of the hidden states (which we denote as $\boldsymbol{x}_t$ at a given time t) of a neural network by equations of the form:

$$\frac{\partial \boldsymbol{x}_t}{\partial t} = D(\boldsymbol{x}_t, \boldsymbol{I}_t, \theta). \tag{1}$$

Where D denotes some kind of model function that estimates the time-derivative of $\boldsymbol{x}_t$. D is a function of $\boldsymbol{x}_t$, which denotes the hidden state of the neuron, $\boldsymbol{I}_t$ which denotes the inputs of the neuron and a learnable parameter vector θ. Updates of the (hidden) state $\boldsymbol{x}_t$ are the computed using some ODE solver which integrates $\frac{\partial \boldsymbol{x}_t}{\partial t}$ over some time-step Δt to compute $\boldsymbol{x}_{t+\Delta t} = \int_t^{t+\Delta t} \frac{\partial \boldsymbol{x}_t}{\partial t} + \boldsymbol{x}_t$.

$$D(\boldsymbol{x}_t, \boldsymbol{I}_t, \theta) = f(\boldsymbol{I}_t, \theta). \tag{2}$$

It is worth noting that in a supervised learning context, this formulation has the advantage of being able to represent irregularly sampled time-sequences. For the control applications we consider here, the sample-rate is imposed by the hardware of our robot and is likely regular. In the original formulation of Chen et al, neural ODEs are not recurrent but a natural extension to recurrent neural networks can be formulated as:

$$D(\boldsymbol{x}_t, \boldsymbol{I}_t, \theta) = f(\boldsymbol{x}_t, \boldsymbol{I}_t, \theta). \tag{3}$$

Where the flow D is not only a function of the input $\boldsymbol{I}_t$ but also of the inner-state $\boldsymbol{x}_t$. Such a model can be though of as a recurrent neural network where the recurrent connection are implemented by an integrator.

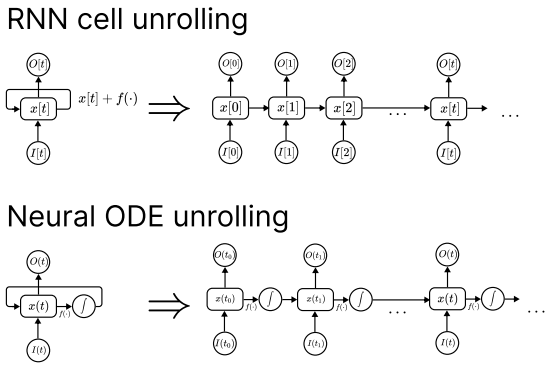

Fig. 1: An illustration of the difference between Neural ODE unrolling v.s. RNN unrolling.

The most straight forward approach to that problem is directly using a neural network to model the flow D, this is the approach chosen by [8] (which is often referred to as "Neural-ODEs"). An alternative provided by an earlier contribution is the so called "Continuous-Time Recurrent Neural Network" model (CT-RNNs) first proposed by Funahashi and Nakamura in [2], which pick D as:

$$D(\boldsymbol{x}_t, \boldsymbol{I}_t, \theta) = -\frac{\boldsymbol{x}_t}{\tau} + f(\boldsymbol{x}_t, \boldsymbol{I}_t, \theta). \tag{4}$$

where τ is a fixed time constant (according to our formalism it is an element of the vector A) and f a non-linear activation function. The fixed time-constant is is introduced to induce a decay in the behavior of the neurons which is meant to enable recurrent connections which avoiding unstable neuron dynamics. In their original paper, Funahashi and Nakamura propose to use $f = \sum_{j=1}^{m} w_{i,j} \cdot \sigma\left(x_{i,t}\right) + I_i(t)$, where the weights wi, j are elements of the vector θ.

More recently, one suggested implementation that seems to give better performance is proposed by Hasani and Lechner though so-called "Liquid Time-Constant networks (LTCs)" [15]. In that case the activation f affects both the time-constant and the non-linearity

(hence make the time-constant "liquid"), this approach corresponds to the following time-derivative model:

$$D(\boldsymbol{x}_t, \boldsymbol{I}_t, \theta) = -\left[\frac{1}{\tau} + A f(\boldsymbol{x}_t, \boldsymbol{I}_t, \theta, A)\right] \boldsymbol{x}_t + f(\boldsymbol{x}_t, \boldsymbol{I}_t, \theta). \quad (5)$$

Where A is a so called *bias* parameter which con-trolls the non-linearity in the synaptic response. In practice, Hasani and Lechner propose to use the following activation function: $f(\boldsymbol{x}_t, \boldsymbol{I}_t, \theta, A) = \tanh(w^{inner}\boldsymbol{x} + w^{inputs}\boldsymbol{I} + \mu)$, which is loosely connected to the non-linearity observed in synaptic dynamics between biological neurons [14].

Table 1 Time-Continuous Neural Network Classes

	Hidden state equation	Recurrent?
CT-RNN	$\frac{\partial \boldsymbol{x}_t}{\partial t} = -\frac{\boldsymbol{x}_t}{\tau} + f(\boldsymbol{x}_t, \boldsymbol{I}_t, \theta)$	Yes
LTC	$\frac{\partial \boldsymbol{x}_t}{\partial t} = -\left[\frac{1}{\tau} + f(\boldsymbol{x}_t, \boldsymbol{I}_t, \theta)\right] \boldsymbol{x}_t + f(\boldsymbol{x}_t, \boldsymbol{I}_t, \theta)$	Yes
Neural-ODE	$\frac{\partial \boldsymbol{x}_t}{\partial t} = f(\boldsymbol{x}_t, \boldsymbol{I}_t, \theta)$	No
RNN-ODE	$\frac{\partial \boldsymbol{x}_t}{\partial t} = f(\boldsymbol{x}_t, \boldsymbol{I}_t, \theta)$	Yes

1.2 Training

Compared to multi-layer perceptrons and RNNs computing gradients on continuous time neural networks is way less obvious because of the integrator step. In the following section we discuss two approaches to the computation of such gradients together with their pros and cons. Two main approaches are possible *backpropagation through time* (BPTT, which is recommanded by Hasani et al. [15]) and the *adjoint sensitivity method* (which is recommended by Chen et al [8]).

Backpropagation through time (BPTT) [15] works by directly computed the gradient of a loss function through the ODE solver (it requires our ODE solver to build a computation graph and then we use *autograd* to compute a gradient). BPTT requires saving the inner state of neurons for every step of the ODE solver, this comes at a high memory cost.

The *adjoint sensitivity method* [8], required saving a so-called adjoint state $\boldsymbol{a}(t) = \frac{\partial L}{\partial x(t)}$ throughout the unrolling of the neural network.

Given some Loss function:

$$L(x(t)) = L\left(x(t_0) + \int_{t_0}^{t} D(x(t), \theta) dt\right),$$

we define our adjoint sensitivity state as $a(t) = \frac{dL}{dx(t)}$ where x is the inner state of our differential equation. The dynamics of $a(t)$ are given by the equation: $\frac{dL(z(t))}{dx(t)}$. The idea here is to use the stored $a(t)$ to compute a gradient:

$$\frac{dL}{d\theta} = -\int_{t_1}^{t_0} a(t)^T \frac{dD(x(t), \boldsymbol{I}_t, \theta)}{d\theta} dt$$

Note that we compute the loss backwards through time (from t_1 to t_0) which is why the negative sign appears in front of the integral. Which we can use to perform gradient descent on our model.

Although Hasani et al. suggest that BPTT reduces numerical error during training [15], it's implementation is quite complex and we decide to stick with the adjoint method for our experiments, as we will discuss in the results section, this does not seem to stop us from learning policies.

1.3 Continuous Time Neural Network Cells

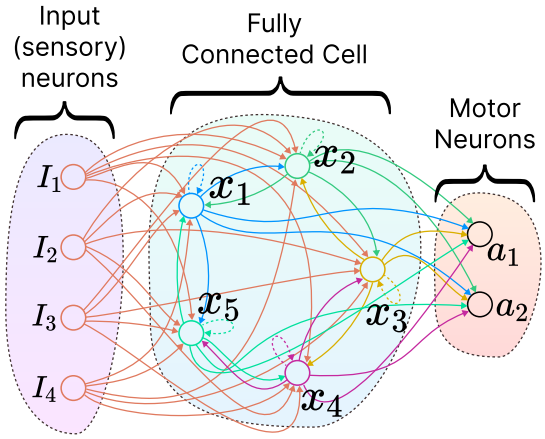

Fig. 2: Fully connected Time-Continuous Neural Network Cell. With 4 inputs ($o_{1,\dots,4}$), 5 inner-neurons (with inner states $x_{1,\dots,5}$) and 2 motor neurons (outputs $I_{1,2}$).

In order to deal with the motor task that we consider in this project we choose to investigate small fully connected neuron cells that take a n-dimensional input, contain k inner-neurons and return d outputs (we call the outputs "*motor neurons*") We provide a visual representation of such a cell in figure 2. Most of our experiments are performed with LTC cells, so we will explicitly derive the forward pass equations for an LTC, but we can build equivalent cells for RNN-ODE and CT-RNN cells in a very similar fashion.

For a given inner-neuron i, the flow of it's state is computed according to the following equation:

$$\frac{dx_i}{dt} = f_i(\mathbf{x}, \mathbf{I}, \theta)$$
$$= -\frac{x_i}{\tau_i} + \tanh\left(\sum_{j=1}^{n} w_{ij}^{inputs} I_j + \sum_{j=1}^{n} w_{ij}^{inner} x_j + \mu_i\right)(A_i - x_i).$$

The states of each neurons are then passed through a linear layer to compute the output values as follows. Each output value's is computed as:

$$o_i = \sum_{j=1}^{k} w_{ij}^{output} x_j$$

Note that this implies that each neuron's flow is influenced by it's input, every single other neuron on the same layer and also itself (it has a recurrent connection). This gives $k \cdot (k + n + d)$ connections in a single LTC cell, which is much denser than a typical multi-layer perceptron. To take a concrete example, in the implementation section when we discuss a 16 neuron LTC cell for solving the cart-pole problem, it has 352 connections for 16 neurons, where a 16 neuron perceptron with a single hidden-layer perceptron would only have 96.

2 Policy Gradient Methods and Reinforcement Learning

In order to optimize the continuous time models we choose to use reinforcement learning methods, the following section covers the Markov Decision Process formalism and gives an introduction to the main ideas behind Proximal Policy Optimization (PPO) [7], the algorithm that we will use to train our model.

2.1 Markov Decision Processes

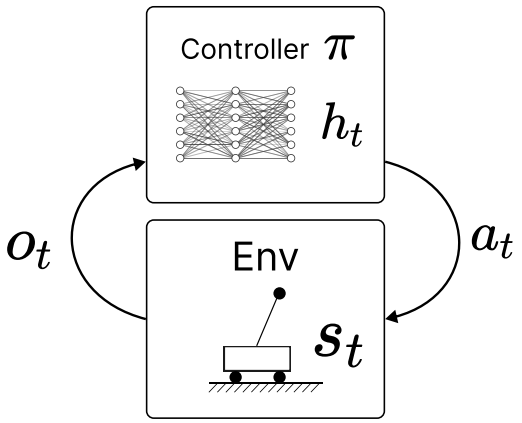

Fig. 3: A visual representation of the different components required to define an RL problem. An environment, which keeps state and defines the transition probability function, and a policy function (which may have a hidden state h) which takes observations o_t of s_t as inputs and returns actions a_t as outputs.

Formulating our control problem as a reinforcement learning problem requires us to write out a control task as a partially observable Markov decision process (POMDP). We define a POMDP through:

1. a (partially observable) state space $\mathcal{S}$ with states $s_t \in \mathbb{R}^n$,
2. an observation space $\mathcal{O}$ with states $o_t \in \mathbb{R}^m$, where observations o_t give some information about the true states s_t,
3. an action space $\mathcal{A}$ with action $a_t \in \mathcal{A} \subseteq \mathbb{R}^d$, which gives the possible actions that the agent can take in a given state,
4. an unknown transition probability function $P(s_{t+1}|s_t, a_t)$, which gives the probability that the system transitions to state s_{t+1} if it is in state s_t and action a_t is taken,
5. a reward function $R : (s_{t+1}, s_t, a_t) \to \mathbb{R}$.

Given a POMDP, a discount factor $\gamma \in [0, 1)$ and a policy function $\pi : \mathcal{S} \to \mathcal{A}$ we can compute the *expected discounted reward* using the Bellman equation:

$$J\pi_\theta = \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_{t+1}, s_t, \pi_\theta(o_t))\right].$$

The problem of reinforcement learning is formulated as the optimization of some parametrizable policy function π (where the parameters are denoted θ) over the POMDP process that ensures the the J value is maximized in expectation across all states, i.e.:

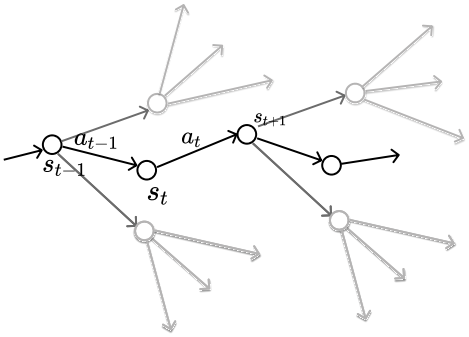

Fig. 4: One can think of a Markov Decision Process as a tree of possible states $s_t, s_{t+1}, ...$ connected by branches associated with the transition probability function $P(s_{t+1}|s_t, a_t)$.

$$\pi^* = \text{argmax}_{\pi_\theta} \mathbb{E}\left[\sum_{t=0}^{\infty} \gamma^t R(s_{t+1}, s_t, \pi_\theta(o_t))\right].$$

2.2 Policy Gradient Methods

We will consider a *policy gradient* reinforcement learning method, such an approach is the most natural for a continuous control task, these method work by directly updating a policy function rather than by computing an estimator of the discounted reward for all possible actions (which is what is done in Q-learning methods). Policy gradient RL methods use update rules of the form:

$$\theta_{k+1} = \theta_k + \alpha \nabla_{\theta_k} J(\pi_{\theta_k})|_{\theta_k}.$$

Here the tricky part of the method resides in deriving an expression for the gradient $\nabla_{\theta_k} J(\pi_{\theta_k})|_{\theta_k}$ of the expected reward J with respect to the parameters θ of the policy π. These methods are often presented in MDP instead of POMDP form but they generalize well to POMDPs. The derivation of policy gradient is performed as follows:

$$\begin{aligned}
\nabla_{\theta_k} J(\pi_{\theta_k})|_{\theta_k} &= \nabla_{\theta_k} \int_\tau P(\tau|\theta) R(\tau) && \text{Expand expectation} \\
&= \int_\tau \nabla_{\theta_k} P(\tau|\theta) R(\tau) && \text{Use linearity} \\
&= \int_\tau P(\tau|\theta) \nabla_{\theta_k} \log P(\tau|\theta) R(\tau) && \text{Log trick} \\
&= \mathop{\mathbb{E}}_{\tau \sim \pi_\theta} \left[\nabla_{\theta_k} \log P(\tau|\theta) R(\tau)\right] && \text{Take the expectations} \\
&= \mathop{\mathbb{E}}_{\tau \sim \pi_\theta} \left[\nabla_{\theta_k} \sum_{t=0}^{T} \log \pi_\theta(a_t|s_t) R(\tau)\right] && \text{Expand over trajectories}
\end{aligned}$$

Where τ denotes the set of all trajectories $s_0, a_0, s_1, a_1, ...$ in the state space under policy π_θ. In practice one can compute an estimate $\hat{g}$ of $\nabla_{\theta_k} J(\pi_{\theta_k})$ by sampling the MDP a sufficiently large number of time. Such an estimator is written out as follows:

$$\hat{g} = \frac{1}{|\mathcal{D}|} \sum_{\tau \in \mathcal{D}} \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) G_t,$$

where G_t is the expected discounted reward over the remaining steps in the episode.

2.3 Advantage Actor Critic

Policy gradient methods derived as we described above tend to lead to unstable learning. This is largely attributable to the fact that the gradient norms are subject to a high variance [3]. The gradient in the variance of $\nabla_{\theta_k} J(\pi_{\theta_k})$ is proportional to the absolute value of the expected discounted reward over the trajectory $R(\tau)$, we can thus reduce the variance of our gradient estimator by subtracting a baseline to it's reward (this doesn't affect the gradients in expectation and thus the algorithm still converges to the same policy), the most common baseline used in that context is the *on-policy value function* $V_{\pi_\theta}(s)$. This means a policy algorithm such as A2C we will have require two separate networks, an actor network (to implement a policy function) and a a critic network (to implement a value estimator) which both need to train simultaneously. To derive the formulation of advantage actor critic we first observe policy gradient implicitely makes use of Q values.

$$\begin{aligned}
\hat{g} &= \mathop{\mathbb{E}}_{\tau} \left[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) G_t \right] \\
&= \mathop{\mathbb{E}}_{s_0, a_0, \dots} \Bigg[[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) G_t] \qquad \text{Observe that:} \\
&\qquad + \mathop{\mathbb{E}}_{s_{t+1}, r_{t+1}, \dots, s_T, r_T} [G_t] \Bigg] \quad \mathbb{E}[G_t] = Q(s_t, a_t) \\
&= \mathop{\mathbb{E}}_{s_0, a_0, \dots} \left[\sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) Q(s_t, a_t) \right]
\end{aligned}$$

Then we subtract the baseline V as follows:

$$\hat{g} = \frac{1}{|\mathcal{D}|} \sum_{\tau \in \mathcal{D}} \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)(G_t - V(s))$$

Using the Bellman Optimality equation $Q(s_t, a_t) = \mathbb{E}[r_{t+1}] + \gamma V(s_{t+1})$ we have that can write out the advantage function as:

$$\begin{aligned}
A(s_{t+1}, s_t, a_t) &= Q(s_t, a_t) - V(s_t, a_t) \\
&\sim r_{t+1} + \gamma V(s_{t+1}) - V(s_t, a_t)
\end{aligned}$$

This approach leads us the the *Advantage Actor Critic Method* (A2C) which computes it's gradients from an advantage function A instead of a direct reward R:

$$\begin{aligned}
\hat{g} &= \frac{1}{|\mathcal{D}|} \sum_{\tau \in \mathcal{D}} \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t) A(s_{t+1}, s_t, a_t) \\
&= \frac{1}{|\mathcal{D}|} \sum_{\tau \in \mathcal{D}} \sum_{t=0}^{T} \nabla_\theta \log \pi_\theta(a_t|s_t)(r_{t+1} + \gamma V(s_{t+1}) - V(s_t))
\end{aligned}$$

In that framework we train two networks side by side, one of them is updated by the approximative gradient $\hat{g}$, and the other is updated by some loss function so that it converges to correct V values.

Algorithm 1: PPO-Clip

Initialize neural networks π_θ and V_ϕ
Set counter $t \leftarrow 0$, observe s_0 from the env
repeat
  forall *workers* k **do**
    Take action $a_t^{(k)} \leftarrow \pi_\theta(s_t^{(k)})$, **observe reward** $r_t^{(k)}$ and new state $s_{t+1}^{(k)}$
    Compute $R_t^{(k)} = r_t^{(k)} + \gamma V_\phi(s_{t+1})$ and the advantage $A_t^{(k)} = R_t^{(k)} - V_\phi(s_t)$
  Update the policy : optimize the surrogate loss $\hat{L}^{\text{CLIP}}(\theta') = \sum_k \gamma^t \text{clip}(r_{\theta'}(s_t, a_t), 1-\epsilon, 1+\epsilon) A_\theta(s_t, a_t)$ with gradient descent for M epochs
  Update the advantage estimator : ϕ by the gradient of $\sum_K (R_t^{(k)} - V_\phi(s_t^{(k)}))^2$
until *convergence*;

2.4 Proximal Policy Optimization

Starting from A2C one can get further improvements in the stability of the convergence by constraining gradient descent to a small region in parameter space for each mini-batch. Algorithms making use of this idea (Trust Region Policy Optimization and Proxy Policy Optimization [6]) are giving state of the art results for continuous control tasks. A really nice walkthrough of this derivation can be found in Johanni Brea's lecture notes [17] for the course Artificial Neural Networks at EPFL.

A rough intuition of why these methods work can be gathered from writing out the expected policy improvements $J(\theta') - J(\Theta)$ of A2C.

$$\begin{aligned}
J(\theta') - J(\theta) &= J(\theta') - \mathop{\mathbb{E}}_{s_0 \sim p(s_0)} [V_\theta(s_0)] && \text{By def. of } J(\theta) \\
&= J(\theta') - \mathop{\mathbb{E}}_{s_t, a_t \sim \theta'} [V_\theta(s_0)] && \text{Take the expectation over all state action pairs (for policy } \pi(\theta')) \\
&= J(\theta') - \mathop{\mathbb{E}}_{s_t, a_t \sim \theta'} \left[\underbrace{\sum_{t=0}^{\infty} \gamma^t V_\theta(s_t) - \sum_{t=1}^{\infty} \gamma^t V_\theta(s_t)}_{\sum_{t=0}^{\infty} \gamma^t [V_\theta(s_t) - \gamma V_\theta(s_t)]} \right] && \text{Write out the expectation as a collapsing sum} \\
&= \mathop{\mathbb{E}}_{s_t, a_t \sim \theta'} \left[\sum_{t=0}^{\infty} \gamma^t \underbrace{[r^a_{s_t \to s_{t+1}} - \gamma V_\theta(s_{t+1}) - V_\theta(s_t)]}_{A_\theta(s_t, a_t)} \right] && \text{Plugging in the definiton of } J(\theta') \\
&= \mathop{\mathbb{E}}_{s_t, a_t \sim \theta'} \left[\sum_{t=0}^{\infty} \gamma^t A_\theta(s_t, a_t) \right] && \text{By definiton of the advantage function} \\
&= \mathop{\mathbb{E}}_{s_t, a_t \sim \theta} \left[\sum_{t=0}^{\infty} \gamma^t \frac{\pi_{\theta'}(s_t, a_t)}{\pi_\theta(s_t, a_t)} A_\theta(s_t, a_t) \right] && \text{Taking expectation over } \pi(\theta) \text{ instead. of } \pi(\theta'))
\end{aligned}$$

We observe that the expected improvement is maximized in expectation when the term $\gamma^t \frac{\pi_{\theta'}(s_t, a_t)}{\pi_\theta(s_t, a_t)} A_\theta(s_t, a_t)$ is maximized. Assuming that the updated policy is close to the previous parametrization (i.e. $\frac{\pi_{\theta'}(s_t, a_t)}{\pi_\theta(s_t, a_t)} \approx 1$). We can directly optimize (maximize) an estimator of the improvement that we call $\hat{L}(\theta')$:

$$\hat{L}(\theta') = \sum_{t=0}^{\infty} \gamma^t \overbrace{\frac{\pi_{\theta'}(s_t, a_t)}{\pi_\theta(s_t, a_t)}}^{= r_{\theta'}(s_t, a_t) \approx 1} A_\theta(s_t, a_t).$$

There are several approaches to keep the updated policy "close" to the previous installment ($r_{\theta'}(s_t, a_t) \approx 1$), one of the is theM compute an estimator of the KL divergence between both policies (which is what is done in TRPO) another is simply clip the loss at some maximum according to the following clipping function:

$$\text{clip}(x, \mu, \nu) = \begin{cases} \nu & \text{if } x < \nu \\ x & \text{if } \nu < x < \mu \\ \mu & \text{if } \mu < x \end{cases}.$$

We called the clipped loss function the "*surrogate loss*". This gives rise to the *PPO-Clip* algorithm for which we give a pseudocode implementation in algorithm 1.

Algorithm 2: Batched PPO-Clip

Initialize neural networks π_θ and V_ϕ
Set counter $t \leftarrow 0$, observe s_0 from the env
repeat
  Initialize a trajectory batch $\tau \leftarrow (\tau^{(1)}, \tau^{(2)}, ..., \tau^{(k)})$
  for $t \in [0, \textit{episode lenght}]$ **do**
    forall *workers* k **do**
      Take action $a_t^{(k)} \leftarrow \pi_\theta(s_t^{(k)})$, **observe reward** $r_t^{(k)}$ and new state $s_{t+1}^{(k)}$
      Compute $R_t^{(k)} = r_t^{(k)} + \gamma V_\phi(s_{t+1})$ and the advantage $A_t^{(k)} = R_t^{(k)} - V_\phi(s_t)$
      Append $s_t^{(k)}, a_t^{(k)}, s_{t+1}^{(k)}, x_t^{(k)}, R_t^{(k)}, A_t^{(k)}$ to the trajectory $\tau^{(t)}$ associated with the worker
  while *trajectory batch* τ *is not empty* **do**
    sample mini-batch from τ (and remove sampled elements)
    Update the policy : optimize the surrogate loss over sampled mini-batch $\hat{L}^{\text{CLIP}}(\theta') = \sum_k \gamma^t \text{clip}(r_{\theta'}(s_t, a_t), 1-\epsilon, 1+\epsilon) A_\theta(s_t, a_t)$ with gradient descent for M epochs
    Update the advantage estimator : ϕ by the gradient of $\sum_K (R_t^{(k)} - V_\phi(s_t^{(}k)))^2$ over sampled mini-batch
until *convergence*;

This reference implementation assumes that the policy is updated as the policy goes through environment steps. In practice this causes trouble when implementing the algorithm as it forces the batch size k to be directly fixed by the number of parallel workers. An alternative approach consists in sampling the environment for several steps (we call this sequence of steps an "*episode*") and then to reconstitute batches after the fact for training. This is what we will implement for our experiments. Algorithm 2 gives pseudocode for such an implementation.

2.5 An architecture for training CTNNs using PPO

Dealing with the continuous time nature of the CTNN cells doesn't cause much trouble when implementing a reinforcement learning algorithm as it mostly just changes the back-propagation algorithm but not the loss or the variables that need to be stored in the process. On the other hand the fact that our CTNN cells are recurrent (have an inner state) changes the process as we somehow need to implement the *unrolling* of the inner states.

The way we do it is by saving the inner states $x_i^{(t)}$ of each neuron at each time step in every trace and then to consider them to be observations when training the policy, this allows for a minimal amount of changes to the PPO architecture at the cost of a relatively high memory use.

3 Implementation of CTNN in a Reinforcement Learning framework

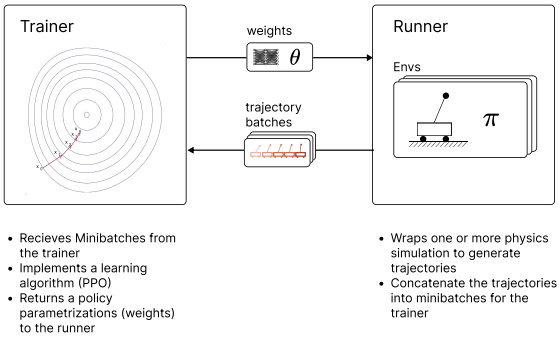

Fig. 5: An illustration of the main components of a reinforcement learning framework. The two main components are left: an implementation of the RL algorithm, and right a wrapper for physics simulations (or environment as they are often called in the RL literature). One of the key functions of such a system is handling the trajectories generated by the environment and converting them into mini-batches, depending on the framework this can be done as part of the runner or trainer components.

The reinforcement learning-based training of a time-continuous neuron cell requires the setup of a learning environment. Such an environment must be able to:

1. simulate the POMDP (since we investigate a continuous control task, this is the physics of our controlled system),
2. implement the policy model (in our case our time-continuous neuron cells) which we discussed in section 1.3,
3. train it using our RL algorithm (PPO) which we discussed discussed in section 2.4.

Because of the complexity of such an environment we make use of a reinforcement learning framework that structures the data collection, the learning and the evaluation of RL policies. The reinforcement learning research community has come up with several such frameworks, notable examples are *stable baselines* [9], *stable baselines 3* [16], *Ray RLLib* [10] and *ACME* [13]. The requirements of our project lead us to seriously consider two main frameworks: *Ray RLLib* [10] and the less known *DERL* [11], which we will now compare in further details.

3.1 Ray RLLib

Our first investigation into the problem we stated were performed within the *Ray RLLib* framework. *Ray* is an open source machine learning framework intended form large-scale distributed computing, its development started at UC Berkley's RISE Lab. *Ray RLLib* [10] is a reinforcement learning framework built on top of *Ray* with the intent of allowing for fast training of RL policies on distributed hardware (for instance on a cluster). Ray features implementation of most state of the art reinforcement learning algorithms (for instance *PPO*, it's asynchronous variant *APPO*, *IMPALA*, various Q-learning derivatives as well as imitation learning algorithms) and bindings with most environments used for continuous control (most notably *PyBullet* and *Mujoco*). *RLLib* allows for implementing policies with either *tensorflow* [5] or *pytorch* [12].

The main argument for the use of *RLLib* is the performance it provides when deployed on a cluster, furthermore much of my direct supervisor Dr. Bellegarda's work on reinforcement learning for the control of complex quadruped robots was performed within this framework, which may open a door to experiments on such systems with less work involved.

Nonetheless I ended up abandoning *RLLib*, this is because of several drawbacks, first most *RLLib*'s implementation either do not support recurrent policies and are often shipped with bugs for recurrent policies [18]. This lack of support of recurrent policies makes the implementation of recurrent continuous time cells such as the ones we discuss in section 1.3, and the fact that *RLLib*'s documentation doesn't clearly specifies which model classes are unsupported further complicates the work. The second big argument for abandoning RLLib is the high complexity of implementing relatively simple changes to algorithms within the framework, this is partly a product of the fact that RLLib applications work as independent processes networked together. This makes the debugging of RLLib applications especially difficult. Furthermore the size of *Ray*'s codebase further participates in slowing the process.

3.2 DERL

DERL [11] is a lightweight reinforcement learning framework written by Mikhail Konobeev, an EPFL Phd student, providing implementations of *PPO*, *SAC*, *A2C* and *DQN* as well as bindings with continuous control environments such as *Mujoco* and *Pybullet*. Unlike *RLLib*, *DERL* isn't intended for distributed computation and it's computation are for the most part run as part of a single python process, which makes debugging much more straightforward. Furthermore the light codebase (as a quick comparison the *Ray* git repository contains 643288 lines of python whereas the *DERL* repository only contains 4314) makes modifications much easier to carry-out. On the other hand the non-distributed nature of it's implementation makes it much slower than *RLLib* and greatly reduces the potential performance benefits that could be hoped for when running it on a cluster. Because of the exploratory nature of the project we deem this trade-off interesting.

The version of *DERL* I used (`06fcd44`) only supports *tensorflow* (and not *pytorch*) so it imposes the choice of the Deep Learning Library. One big argument for the use of *DERL* was an example implementation of non-recurrent neural ODEs for control [11], this provides a starting point for building of continuous time neural network cells. *DERL* doesn't provide support for recurrent policies, so a large part of my work resided in the implementation of training with recurrent policies, which mostly consists in handling the saving of the inner state of recurrent model as part of the trajectories.

3.3 Mujoco

The two most used physics simulator for reinforcement learning are *Pybullet* [19] and *Mujoco* [4]. The quadruped simulations used in *BIOROB* are most often built within the *Pybullet* environment which until recently was the best open-source choice for the task. This changed as *Google Deep Mind* bought and open-sourced the - until then proprietary - *Mujoco* physics engine in may 2022. *Mujoco* provides a slightly faster implementation of a physics simulation and as *DERL* provides bindings to the *Mujoco* gym environments developed at *OpenAI* it made sense to make use of the *Mujoco* environment in the context of the project.

4 Results and Discussion

4.1 Experimental Setup

In order to investigate the efficiency of CTNN cells we setup a simple (and rather classical in the RL literature [1]) motor control experiment: the cart-pole problem. We use the *Mujoco* gym implementation of cart-pole together with our *DERL* implementation of *PPO* training for CTNN policies to perform our experiments. The code used for the experiments is publicly available at https://github.com/RenardDesNeiges/CTNN_Policies_DERL.

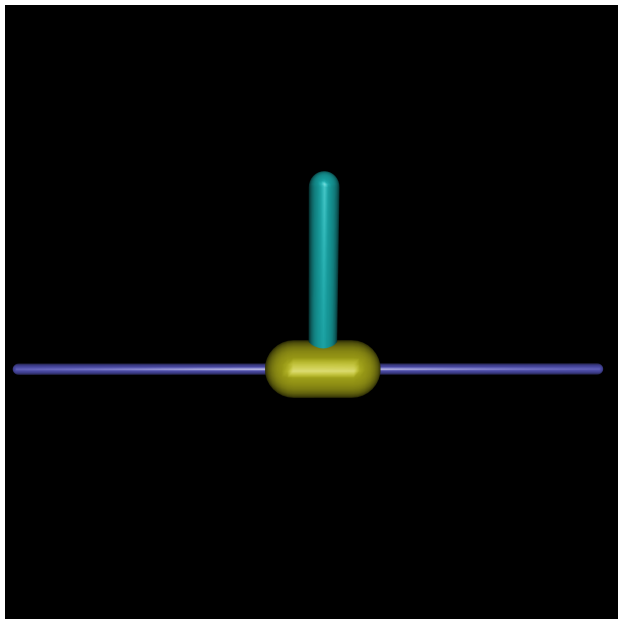

Fig. 6: The cart-pole system, as rendered inside of *Mujoco*'s built-in viewer.

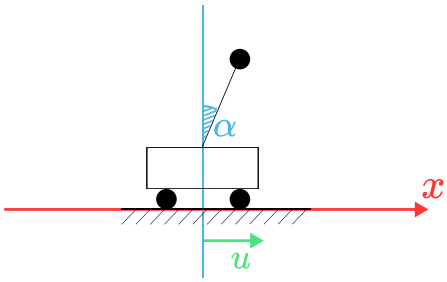

Fig. 7: The cart-pole system, with both coordinates x and α, as well as the control variable u (a linear force on the cart) clearly labeled.

We focused our experiments on *LTC* cells with various neuron counts as an investigation. For all of our cart-pole experiments we use the following reward:

$$r_t = \begin{cases} -2 & \text{if } |\alpha| > \alpha_{\max} \text{ (failure)} \\ 1 - qu^2 & \text{otherwise} \end{cases}$$

Where $q = 0.2$ is a penalty on control (where u is a force measured in newtons) and where the angle $\alpha_{\max} = 0.209$ is used to define failure (with α the pole-angle measured in radiant).

We find that the LTC cell trained with PPO can find efficient solutions to the cart-pole problem, but that it does so with a particularly low sample efficiency (compared to a reference *MLP* model trained with a perceptron).

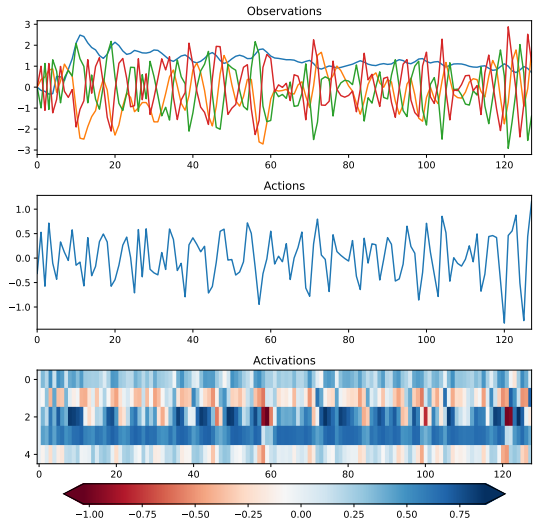

Fig. 8: Visualization of (from the top down) observation, action and neuron activations together for a 5 neuron LTC cell trained with PPO on a cart-pole *Mujoco* environment. The activation plot is structured as follows, each row correspond to a single neuron is colored along the t horizontal axis according to the activation x_i of that neuron at time t.

4.2 Neuron Count Experiments

We run experiments with various neuron counts $(5, 6, 7, 8, 16,$ and 32 neurons). The main questions we aim to answer with such a protocol are "*Is there a minimum neuron count to solve such a problem?*" (and how different is it from the minimum neuron count with a perceptron) and "*how does the neuron count affect the quality of solution?*".

4.3 Sample efficiency

TODO : Influence of nc and mask?

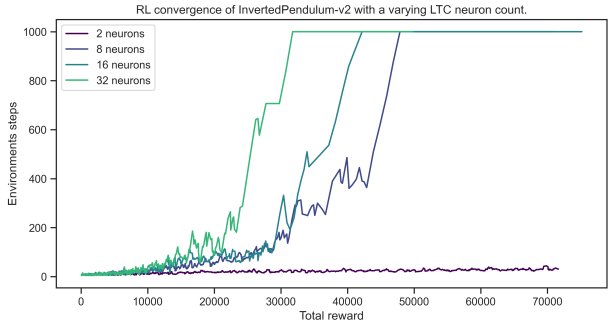

Fig. 9

4.4 Resilience to noise?

TODO : Influence of nc and mask?

4.5 Resilience to noise?

TODO : Influence of nc and mask?

4.6 Discussion of the results

TODO : Discuss the usability of the learned policies, what can we infer from this

4.7 Further work section

TODO : Discuss what I didn't have time to do?

References

[1] Andrew G. Barto, Richard S. Sutton, and Charles W. Anderson. "Neuronlike adaptive elements that can solve difficult learning control problems". In: *IEEE Transactions on Systems, Man, and Cybernetics* SMC-13.5 (1983), pp. 834–846. DOI: 10.1109/TSMC.1983.6313077.

[2] Ken-ichi Funahashi and Yuichi Nakamura. "Approximation of dynamical systems by continuous time recurrent neural networks". In: *Neural Networks* 6 (1993), pp. 801–806.

[3] Vijay Konda and John Tsitsiklis. "Actor-Critic Algorithms". In: *SIAM Journal on Control and Optimization*. MIT Press, 2000, pp. 1008–1014.

[4] Emanuel Todorov, Tom Erez, and Yuval Tassa. "MuJoCo: A physics engine for model-based control." In: *IROS*. IEEE, 2012, pp. 5026–5033. ISBN: 978-1-4673-1737-5. URL: http://dblp.uni-trier.de/db/conf/iros/iros2012.html#TodorovET12.

[5] Martín Abadi et al. *TensorFlow: Large-Scale Machine Learning on Heterogeneous Systems*. Software available from tensorflow.org. 2015. URL: https://www.tensorflow.org/.

[6] John Schulman et al. "Proximal Policy Optimization Algorithms". In: *ArXiv* abs/1707.06347 (2017).

[7] John Schulman et al. "Proximal Policy Optimization Algorithms." In: *CoRR* abs/1707.06347 (2017). URL: http://dblp.uni-trier.de/db/journals/corr/corr1707.html#SchulmanWDRK17.

[8] Tian Qi Chen et al. "Neural Ordinary Differential Equations". In: *NeurIPS*. 2018.

[9] Ashley Hill et al. *Stable Baselines*. https://github.com/hill-a/stable-baselines. 2018.

[10] Eric Liang et al. "RLlib: Abstractions for Distributed Reinforcement Learning". In: *Proceedings of the 35th International Conference on Machine Learning*. Ed. by Jennifer Dy and Andreas Krause. Vol. 80. Proceedings of Machine Learning Research. PMLR, Oct. 2018, pp. 3053–3062. URL: https://proceedings.mlr.press/v80/liang18b.html.

[11] Mikhail Konobeev. *Neural Ordinary Differential Equations for Continuous Control*. https://github.com/MichaelKonobeev/neuralode-rl. 2019.

[12] Adam Paszke et al. "PyTorch: An Imperative Style, High-Performance Deep Learning Library". In: *Advances in Neural Information Processing Systems 32*. Curran Associates, Inc., 2019, pp. 8024–8035. URL: http://papers.neurips.cc/paper/9015-pytorch-an-imperative-style-high-performance-deep-learning-library.pdf.

[13] Matt Hoffman et al. "Acme: A Research Framework for Distributed Reinforcement Learning". In: *arXiv preprint arXiv:2006.00979* (2020). URL: https://arxiv.org/abs/2006.00979.

[14] Mathias Lechner et al. "Neural circuit policies enabling auditable autonomy". In: *Nature Machine Intelligence* 2 (2020), pp. 642–652.

[15] Ramin M. Hasani et al. "Liquid Time-constant Networks". In: *AAAI*. 2021.

[16] Antonin Raffin et al. "Stable-Baselines3: Reliable Reinforcement Learning Implementations". In: *Journal of Machine Learning Research* 22.268 (2021), pp. 1–8. URL: http://jmlr.org/papers/v22/20-1364.html.

[17] Johanni Brea. *Lecture notes for the course Artificial Neural Networks*. 2022. URL: https://moodle.epfl.ch/pluginfile.php/2779866/mod_resource/content/3/deeprl.pdf.

[18]*Github Issue [Bug] [rllib] RNN sequencing is incorrect*. https://github.com/ray-project/ray/issues/19976. Accessed: 2022-06-09.
[19]Erwin Coumans and Yunfei Bai. *PyBullet, a Python module for physics simulation for games, robotics and machine learning*. http://pybullet.org. 2016–2021.