

Fido: A Universal Robot Control System Using Reinforcement Learning with Limited Feedback

Joshua Gruenstein Michael Truell

Control System Objectives

Fido was created to fulfill the following goals:

- **Trainability:** Allow both human and autonomous training and retraining rather than programming
- **Universality:** Run on any robot and perform any task, even without prior knowledge of the host
- **Performance:** Require few learning iterations and low latency for quick and efficient training

To achieve this, we developed a novel machine learning algorithm utilizing wire-fitted Q-Learning with a dynamically resized neural network, a confidence-based probabilistic action selection model, and experience replay. The control system could be trained at a rate **4 times** the industry standard, allowing practicality over traditional pre-programmed control systems.

System Overview

Fido Control System Algorithm

- 1: Initialize empty replay memory M
- 2: Initialize neural network with random weights
- 3: Uncertainty value $u \leftarrow \infty$
- 4: **loop**
- 5: Observe state s_t
- 6: Feed vector s_t to model and generate continuous function $Q(a)$ of action versus reward
- 7: Select action a_t with expected reward r_t^* using a Softmax action selection policy with temperature $T \propto u$ and the generated $Q(a)$ function.
- 8: Execute action a_t and observe reward r_t and new state s_{t+1}
- 9: Store (s_t, a_t, r_t, s_{t+1}) in M
- 10: $u \propto (r_t - r_t^*)^2$
- 11: Sample $n \propto 1/u$ experiences e giving weight to newer experiences.
- 12: Perform a search on model architectures with cost function as error in predicting e when trained on e using stochastic gradient descent and Adadelata
- 13: Update current model with result of search

From a macro perspective, Fido can be viewed as a “black box” system, where a set of inputs (such as sensors) go in and a set of outputs (actions) go out. Fido’s goal is to maximize given reward by altering its behavior. The system operates through the following steps:

1. Sensor values are fed to a neural network.
2. The neural network outputs data points, each is an action and its expected reward.
3. A wire-fitted moving least squares interpolator creates a continuous function of action to expected reward using these data points.
4. An action is chosen using an Softmax selection policy that dynamically adjusts Fido’s exploration level as its uncertainty level changes.
5. After receiving reward, the neural network is trained to output a new set of data points using Adadelata for gradient descent. Dynamic sizing of the neural network and experience replay are employed to improve training time.

Reinforcement Learning

- Fido is a **reinforcement learning system**. Its goal is to accurately estimate the reward for any action it can perform.
- The system uses **Q-Learning**, a popular reinforcement learning algorithm that develops a function that intakes a state-action pair and outputs expected utility.
- Fido utilizes a **neural network** coupled with a **wire-fitted moving least squares interpolator** instead of the traditional table to estimate its Q-function such that relations can be made between similar actions and similar states. Our neural network accepts the state and outputs a few actions along with their expected rewards. Fido interpolates these action-reward pairs to create a continuous function of action to reward.
- Once Fido performs an action and receives a reward (a single **learning iteration**), the control system calculates its **uncertainty value** as the difference between its expected and received reward. Higher uncertainty means that Fido is still learning or is undergoing retraining.
- Fido selects actions to perform using a probabilistic Softmax policy which gives greater weight to actions with greater expected utility, but allows for exploration.
 - Fido dynamically adjusts its exploration level to be proportional to its uncertainty value.
 - This allows Fido to continuously execute the best actions once it has been trained and is sure of its actions, but also allows it to explore when being retrained. This results in lower training times and higher average reward.

Training

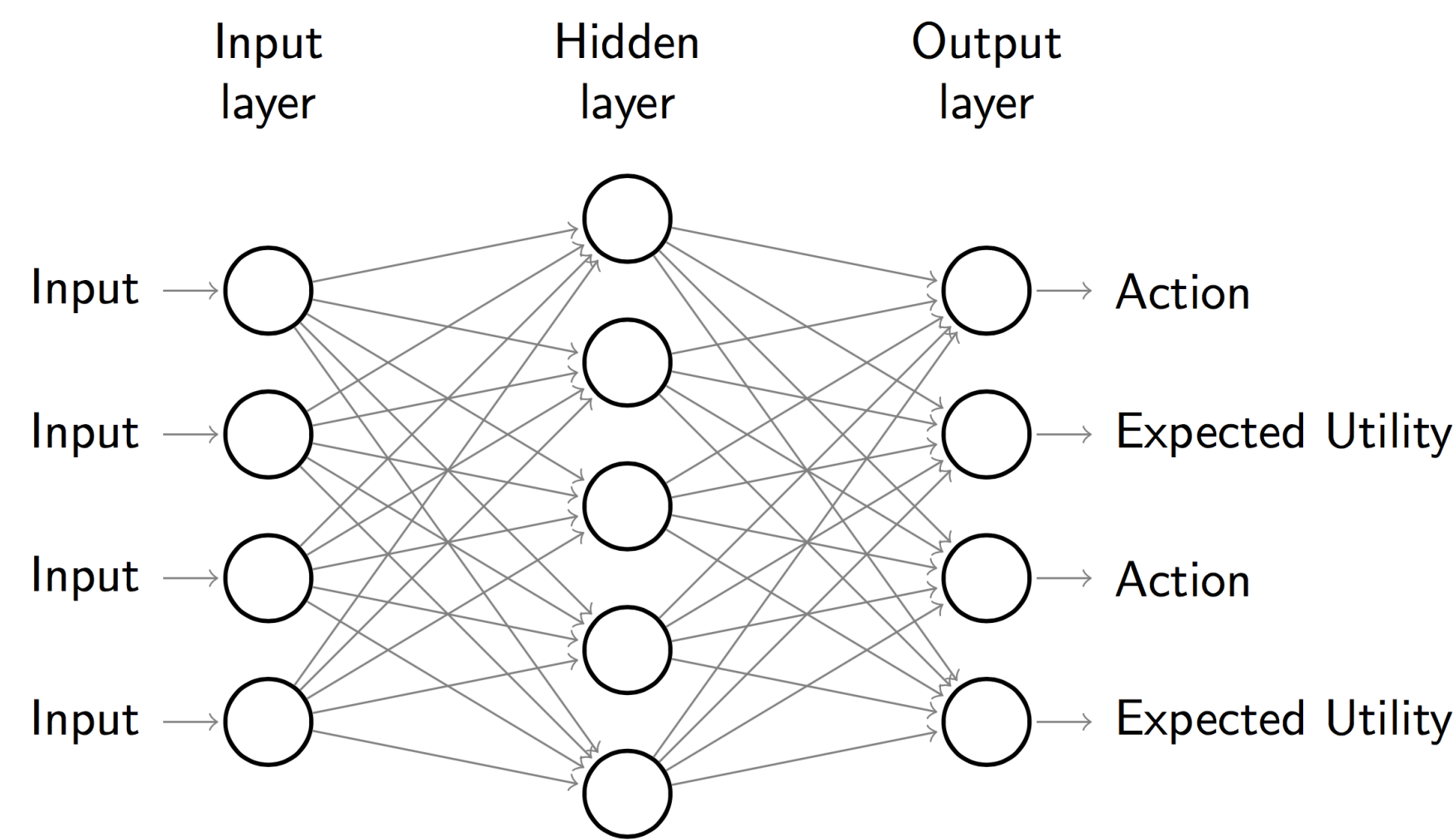


Figure: Fido’s Neural Network

- After every reward iteration, Fido adjusts its model to correct the disparity between its expected reward and the reward it just recieved using **Stochastic Gradient Descent** to modify the action-reward pairs that are outputted by its neural network.
- Fido trains its neural network using the Adadelata training method, eliminating the need to specify a hard-coded learning rate.
- Fido uses **experience replay**, the practice of training a system on past states, actions, and rewards to decrease training time. Fido gives weight to newer experiences. If uncertainty value of the model begins to rise, Fido samples fewer experiences, because past rewards will be incorrect if Fido is being trained on a new task.
- After each reward iteration, Fido efficiently grows and prunes its neural network to the optimal size and architecture for the task at hand, using neuron sensitivity approximations.

Results

Results were gathered both from simulation and hardware for a variety of tasks. Gathered data for each task included the number of learning iterations, average action selection time, and average training time (latency in updating Fido’s model).

Table: Fido Results in Simulation (400 trials per task)

Task	Learning Iterations	Action Selection (ms)	Training Time (ms)
Flash	3	0.	28
Float to Point	14	1	63
Drive to Point	16	1	94
Line Following	10	0.	90
Noisy Line Following	12	0.	323

Table: Fido Results on Thing One (20 trials per task)

Task	Learning Iterations	Action Selection (ms)	Training Time (ms)
Stay Still	3	1	80
Drive to Point	18	4	307

Table: Fido Results on Thing Two (20 trials per task)

Task	Learning Iterations	Action Selection (ms)	Training Time (ms)
Drive Straight	13	2	215
Line Following	15	21	398
Fetch	8	1	246
Limping Line Following	6	20	464

Table: Fido Results on Thing Three (20 trials per task)

Task	Learning Iterations	Action Selection (ms)	Training Time (ms)
Draw Square	5	5	40
Ping Pong	8	3	379

Fido vs. Industry Standard

Fido’s performance in completing various simulator tasks was tested against the industry standard of discrete Q-Learning using a neural network. On average, Fido required **over four times fewer** learning iterations than discrete Q-Learning for a given task.

Table: Fido Results Compared to Discrete Q-Learning

Task	Fido Learning Iterations	Discrete Learning Iterations
Flash	3	16
Float to Point	14	56
Drive to Point	16	69
Line Following	10	42

Implementation

The Fido control system was programmed in C++, with no external dependencies. Three hardware implementations (named Things One, Two, and Three) and a simulator were constructed to test Fido’s performance. The first robot utilized a differential drive system and was powered by an Intel Edison, the second used a holonomic wheel arrangement and ran on a \$5 Raspberry Pi Zero, while the third was a 4-axis arm controlled by a Raspberry Pi Zero. The “things” were administered reward over Wi-Fi from a cross-platform mobile app.

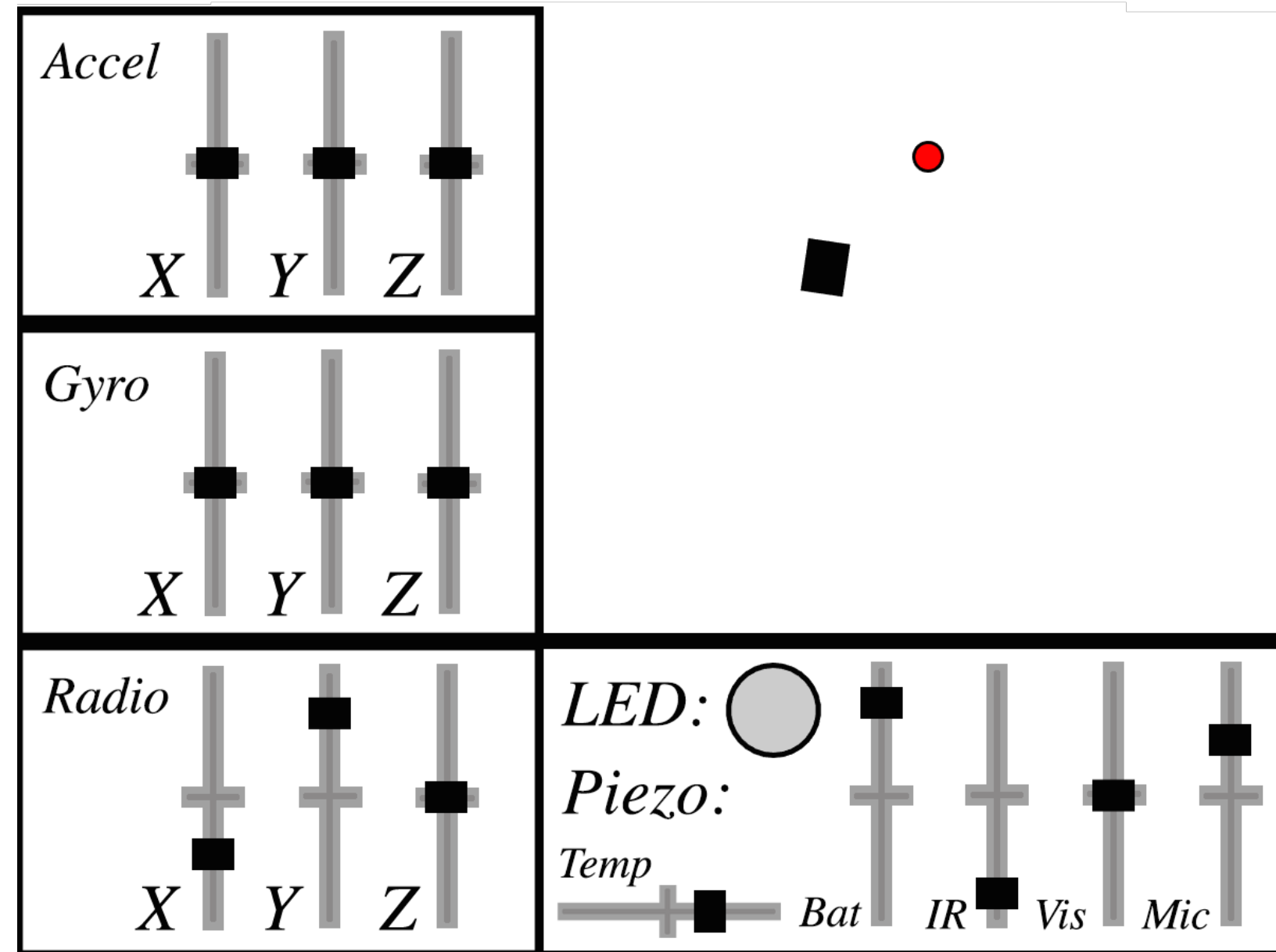


Figure: Fido Simulator Graphical User Interface

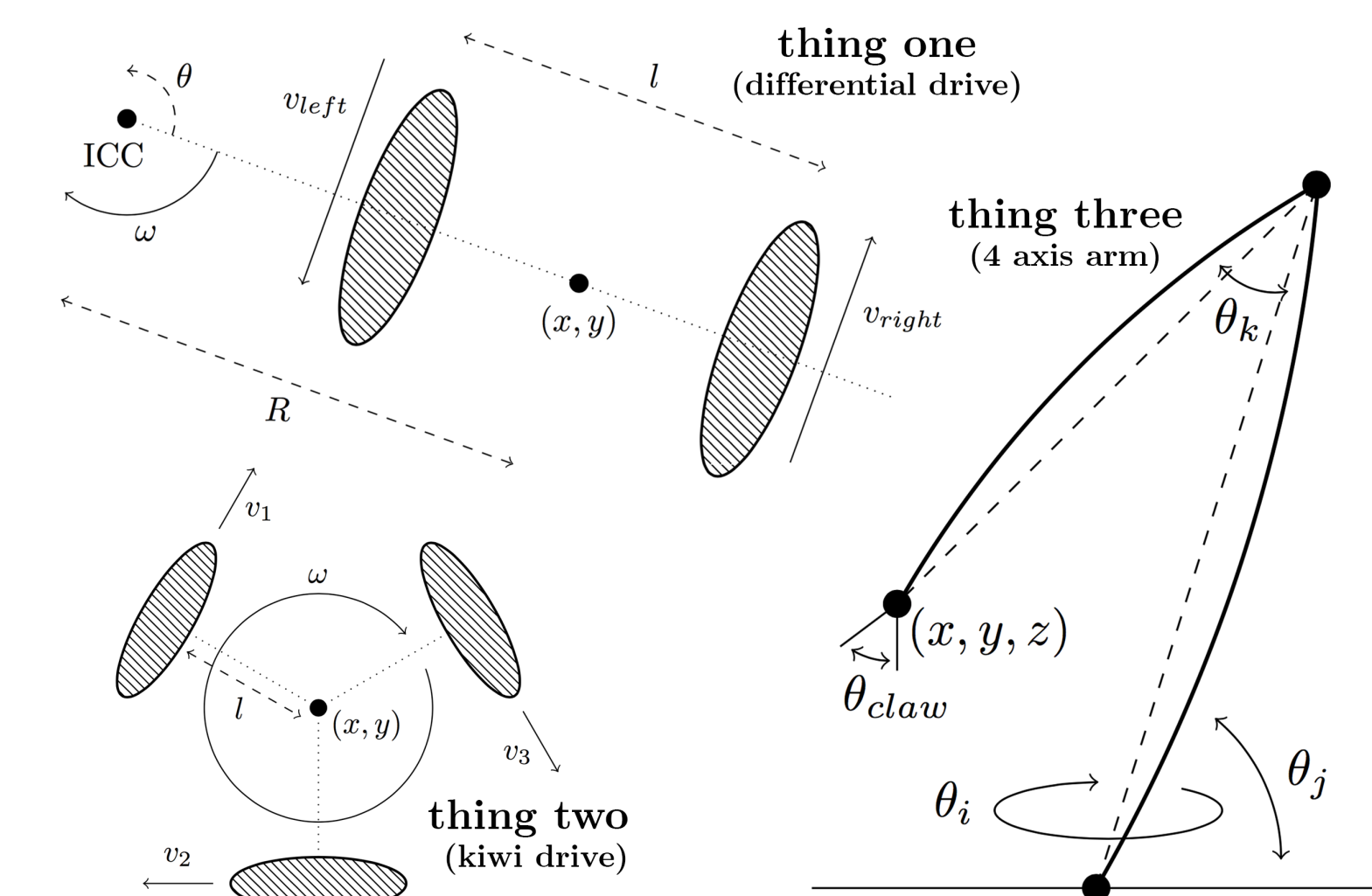


Figure: Fido Hardware Implementation Kinematics

Applications

Fido provides numerous advantages over traditional procedural-programmed control systems, making it practical for a number of real world applications. As Fido can be trained and retrained without expert knowledge, it could be useful in making robotics more accessible to consumers and small-buisness owners. Fido can also be retrained to adapt to new and unexpected circumstances such as malfunctions in the field, an advantageous feature for military applications.