Technical Abstract

Author Alex Darch
Supervisor Dr. Glenn Vinnicombe
Assessor Dr. Ioannis Lestas
St John's College, Cambridge

AlphaZero is a Reinforcement Learning algorithm developed by Google Deepmind. It can be interpreted as a Robust Model Predictive Control (RMPC) algorithm that performs a directed tree search through the future states to return a control action rather than optimising over all future states. This allows it to dramatically reduce computational time, whilst searching more likely trajectories to a greater depth. Additionally, AlphaZero is a learning algorithm that trains a neural network representing a policy through self play. The adversary in self play can be interpreted as a sequence of optimally adverse disturbances.

Controlling dynamical systems is an active area of interest and this project looks at the application of AlphaZero to these as a novel method of control. This has a number of potential benefits above traditional optimal control such as the ability to robustly handle disturbances; readily model non-linear systems; and, due to the generality of the algorithm, the ability to be applied to many control problems. The inverted pendulum was chosen as a simple starting dynamical system which also met the requirements of being sufficiently complex such that a trivial solution was not possible.

In order to adapt AlphaZero into a control algorithm, methods from the field of Model Predictive Control (MPC) and Optimal Control have been modified. These include the design of a cost function and the introduction of a type of receding horizon for the state value, although this is retrospectively calculated. Adding these allowed for the algorithm to deal with continuous states and indefinite horizons with no easily discernable "winner", unlike the board games played by AlphaZero. In addition to this, two different methods of state representation are considered: the true state, $\boldsymbol{x}^T = [x,\ \dot{x},\ \theta,\ \dot{\theta}]^T$, and a 2-dimensional (2D) state, $\boldsymbol{x}^{(2D)}$, which histogrammed (binned) the evolution of positions, (x_t, θ_t), over time on a grid. The true state has the benefit of being continuous with a low number of dimensions, which made it more ideal for this specific problem. However, the 2D state does not rely on measurements of the system velocities. Relying purely on positional information allows for a more generalisable state space representation as, in real systems, states are often unobservable. Furthermore, the binning of the states was proven to be lossless in the limit of infinite time, assuming a deterministic model and that the neural network could emulate an optimal Kalman Filter.

One of the greatest difficulties in adapting AlphaZero was "power matching" - balancing the effect of each agents actions such that they have an equal effect on the system.

The adversary was chosen to act at the tip of the pendulum so that the adversary would be forced to learn a different policy to the player. However, due to the effect of gravity and a different point of action, the adversary and player had unequal effects on the system. It was found that simply choosing the force of the adversary would either lead to the pendulum falling very quickly, meaning that minimal training examples were recorded; or the adversary was too weak to push over the pendulum. In the case of the adversary being too weak, the adversary invariably learnt to push solely in one direction. This is postulated to be the most efficient way for the adversary to maximise the cost of the system. Due to the adversary learning this, robustness to general disturbances was severely hindered. It was found that the player's internal model of the adversary could not predict the actions of more general adversaries, and therefore the player could not control these.

It has been shown that MCTS can be used to improve the control of a system to great effect, and system non-linearities do not impede this. This method can model any system given that it is deterministic and the player has a perfect knowledge of the action space of the adversary. However, more work is needed for this method to be considered "robust control".