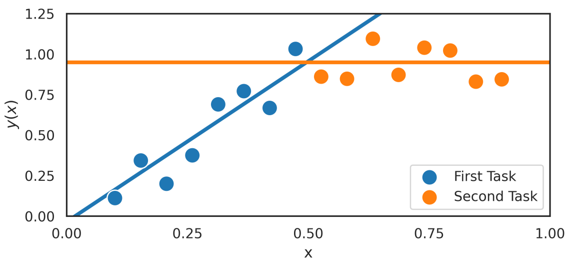
ATLAS: Sparse, Efficient, Spline-Based ANNs with Robustness to Catastrophic Forgetting

ATLAS: Efficient Learning Without Catastrophic Forgetting

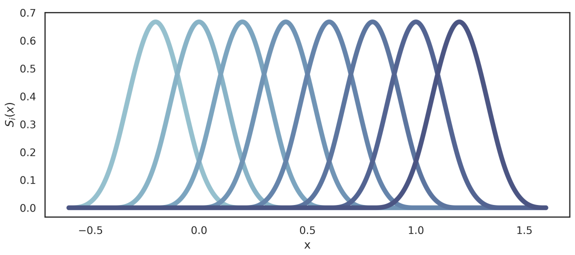
Heinrich van Deventer, Anna Bosman

1 Introduction

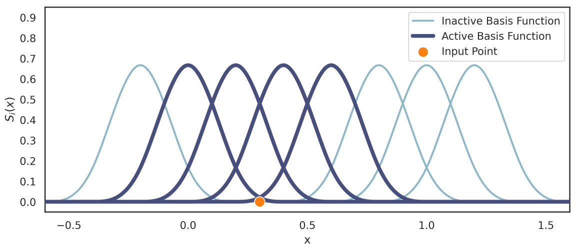
Catastrophic forgetting hinders sequential and continual learning, and can make training slower and less efficient. Even linear models are susceptible to catastrophic forgetting.



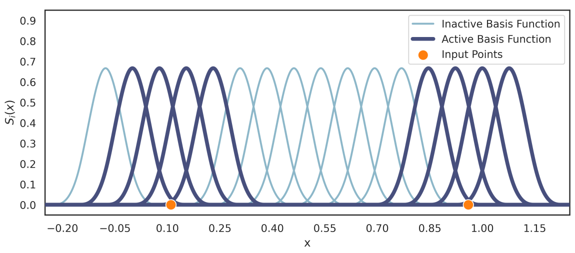
- Globally shared parameters make models susceptible to catastrophic forgetting.
- Piece-wise defined functions do not share parameters over all inputs.
- Cubic B-splines are robust to forgetting.



- Uniform splines have the same shape, and are implemented with an activation function by scaling and translating inputs correctly.



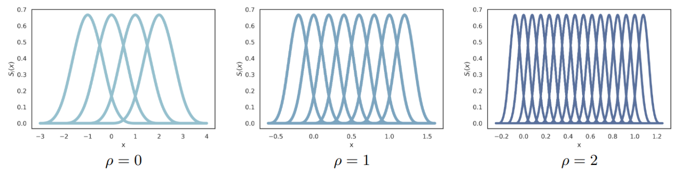
- Only **four** basis functions are **non-zero**, regardless of the number of basis functions.



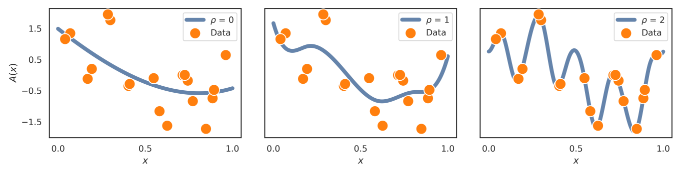
- If two inputs are far enough from each other, then they do not interfere with each other.

2 Single-Variable Functions

Create more powerful function approximators with a larger density of basis functions that are doubled.



- Create minimal model.
- Train model to convergence.
- Increase model size and train again.



3 Universal Function Approximation

Named for carrying the weight of all it must remember, ATLAS is a function approximator of n variables, with mixed-density B-spline functions $f_j(x_j)$, $g_{i,j}(x_j)$, and $h_{i,j}(x_j)$ in the form:

$$A(\vec{x}) := \sum_{j=1}^{n} f_j(x_j) + \sum_{k=1}^{M} \frac{1}{k^2} \exp\left(\Sigma_{j=1}^{n} g_{k,j}(x_j)\right) - \frac{1}{k^2} \exp\left(\Sigma_{j=1}^{n} h_{k,j}(x_j)\right)$$

4 Distal Orthogonality

If two vector inputs differ from each other in each input variable, then the gradient updates are orthogonal. For any $\vec{\mathbf{x}}, \vec{\mathbf{y}} \in D(A) \subset R^n$ and ATLAS model $A(\vec{\mathbf{x}})$ bounded trainable parameters θ_i, there exists a $\delta > 0$ such that:

$$|x_j - y_j| > \delta \ \forall j \in \mathbb{N} \implies \langle \vec{\nabla}_{\vec{\theta}} A(\vec{\mathbf{x}}), \vec{\nabla}_{\vec{\theta}} A(\vec{\mathbf{y}}) \rangle = 0$$

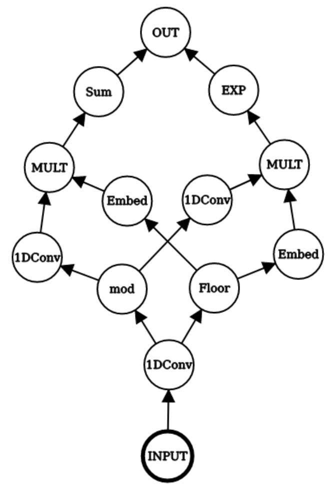
5 Gradient Flow Attenuation

For any $\vec{\mathbf{x}} \in D(A) \subset R^n$ and bounded trainable parameters Θ: if all the mixed-density B-spline functions are bounded, then the gradient vector of trainable parameters for ATLAS is bounded:

$$\left\| \vec{\nabla}_{\vec{\theta}} A(\vec{\mathbf{x}}) \right\|_1 = \sum_{\theta_i \in \Theta} \left| \frac{\partial A}{\partial \theta_i}(\vec{\mathbf{x}}) \right| < U$$

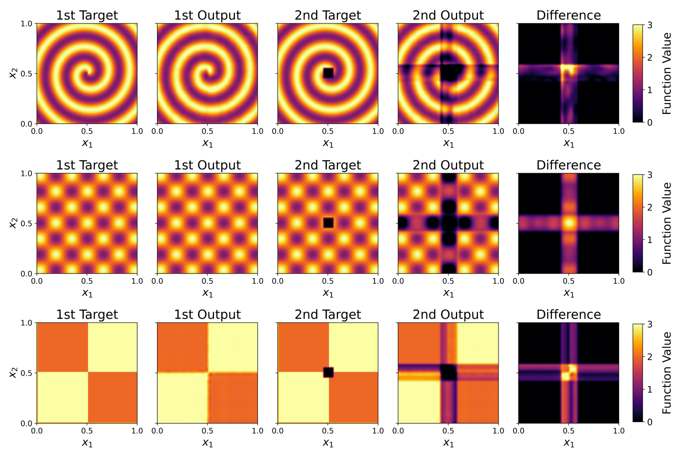
6 Technical Overview

We created an efficient implementation of ATLAS with convolutional layers and embedding layers to look up parameters. Very few redundant computations are made. Only non-zero basis functions are evaluated. The condensed computational graph of ATLAS using 1D convolution, embedding, multiply, and activation layers:



Computational time complexity: $\mathcal{O}(Mn \log \lambda)$, and space complexity: $\mathcal{O}(Mn\lambda)$. At most 2λ basis functions for each single-variable function.

7 Results



- **Theoretical** advances in function approximation and mitigating catastrophic forgetting.
- **Technical** success in developing efficient TensorFlow implementations of ATLAS.
- **Empirical** evidence of memory retention and robustness to catastrophic forgetting.

UNIVERSITEIT VAN PRETORIA
UNIVERSITY OF PRETORIA
YUNIBESITHI YA PRETORIA

Faculty of Engineering, Built Environment and Information Technology

Fakulteit Ingenieurswese, Bou-omgewing en Inligtingtegnologie / Lefapha la Boetšenere, Tikologo ya Kago le Theknolotši ya Tshedimošo

Download the paper →