Universal Approximation in Action: A Lightweight Demo with a Two–Layer ReLU MLP

Author: Aayush Bajaj (abaj.ai)
Date: 6 July 2025
Repo: `README.gif` generator (Python)

Goal. Visually substantiate the *Universal Approximation Theorem* by training a small multi-layer perceptron (MLP) to learn nine qualitatively different functions $f : [-1, 1]^2 \to \mathbb{R}$ (**??**). The resulting GIF cycles through each target surface while showing the network's prediction $\widehat{f}_t$ at training step t.

Architecture. The network is

$$(x, y) \xrightarrow{\text{Linear}} \mathbb{R}^{64} \xrightarrow{\text{ReLU}} \mathbb{R}^{64} \xrightarrow{\text{Linear}} \mathbb{R}^{64} \xrightarrow{\text{ReLU}} \mathbb{R}^{64} \xrightarrow{\text{Linear}} \widehat{f}(x, y) \in \mathbb{R}.$$

With two hidden layers the total parameter count is $\approx 4 \times 10^3$, well below GPU-scale yet sufficient for expressive power.

Training data. Each surface is sampled on a fixed Cartesian grid $\mathcal{X} = \{(x_i, y_j)\}_{i,j=1}^{50} \subset [-1, 1]^2$ (2 500 points). The same grid is re-used for every function.

Loss function. Mean–squared error (MSE) is minimised:

$$\mathcal{L}_t = \frac{1}{|\mathcal{X}|} \sum_{(x,y)\in\mathcal{X}} \big(\widehat{f}_t(x, y) - f(x, y)\big)^2. \tag{1}$$

MSE is chosen for its convexity in the output layer parameters, smooth gradients and compatibility with the visual metric "prediction surface $\approx$ truth". Adam ($\alpha = 2 \times 10^{-2}$) drives optimisation for 400 epochs per function.

Why MSE? Alternatives like binary cross-entropy would be appropriate for *thresholded* glyphs but yield vanishing gradients on large flat regions. MSE penalises *distance* rather than *mis-classification*, keeping the signal informative even where $f \in \{0, 1\}$.

Looping GIF. Every $N = 25$ back-prop steps we capture a frame: (i) network stats, (ii) ground-truth mesh, (iii) current prediction. All frames share an identical 768×256 canvas so ImageIO can stack them:`loop=0` for infinite replay.

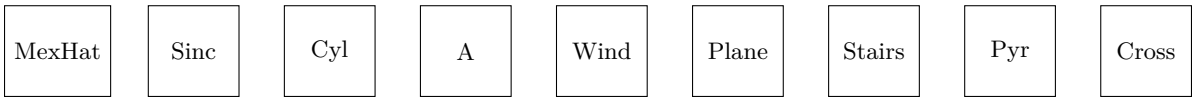

Figure 1: The nine target surfaces—oscillatory, piecewise-constant and polygonal alike—demonstrate depth-2 ReLU universality.

Observations.

- Smooth functions (Mexican Hat, Sinc) fit in < 40 epochs; the piecewise -constant glyphs require ~ 200 epochs due to corner singularities.
- The cylinder's steep wall highlights the ReLU's piecewise-linear nature: the MLP forms concentric linear bands that sharpen with depth.
- Despite identical hyper-parameters for all tasks, the network converges without explicit scheduling—underscoring the robustness of Adam.
