# LassoNet: A Neural Network with Feature Sparsity

Ismael Lemhadri†, Feng Ruan‡, Robert Tibshirani†

†Stanford University, ‡University of California, Berkeley

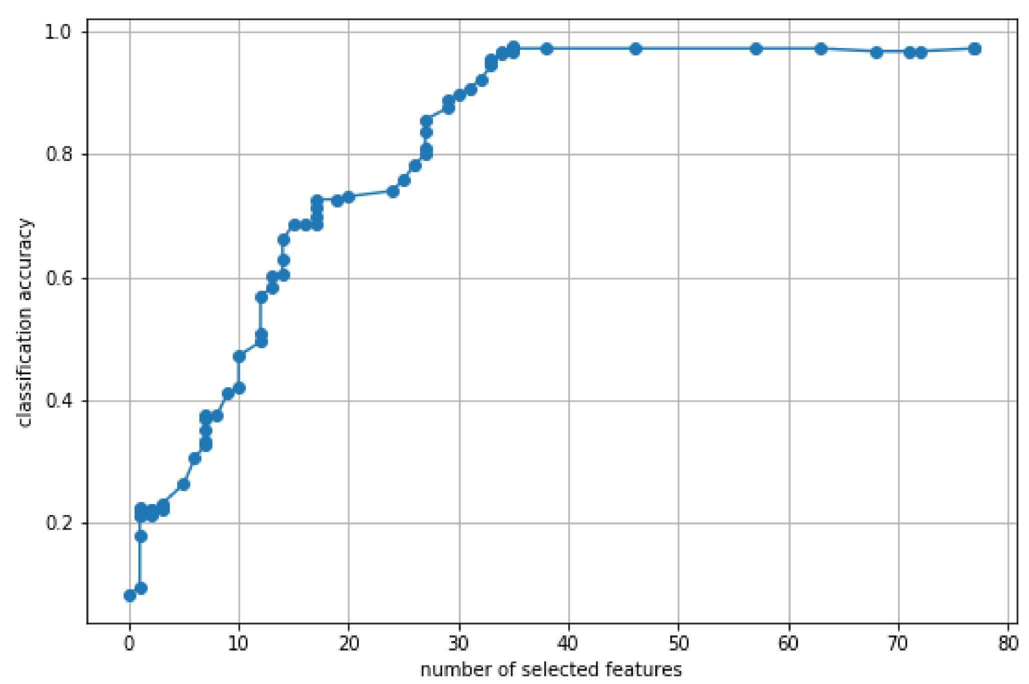
tinyurl.com/lassonet

## Summary

- Much work has been done recently to make neural networks more **interpretable**, and one approach is to arrange for the network to use only a subset of the available features.
- We introduce LassoNet, a neural network framework with global feature selection.
- Our approach enforces a hierarchy: specifically a feature can participate in a hidden unit only if its linear representative is active.
- Unlike other approaches to feature selection for neural nets, our method delivers an entire path of solutions with a range of feature sparsity.
- The method uses projected proximal gradient descent, and generalizes directly to deep networks.
- It can be implemented by adding just a few lines of code to a standard neural network.

## Motivating example

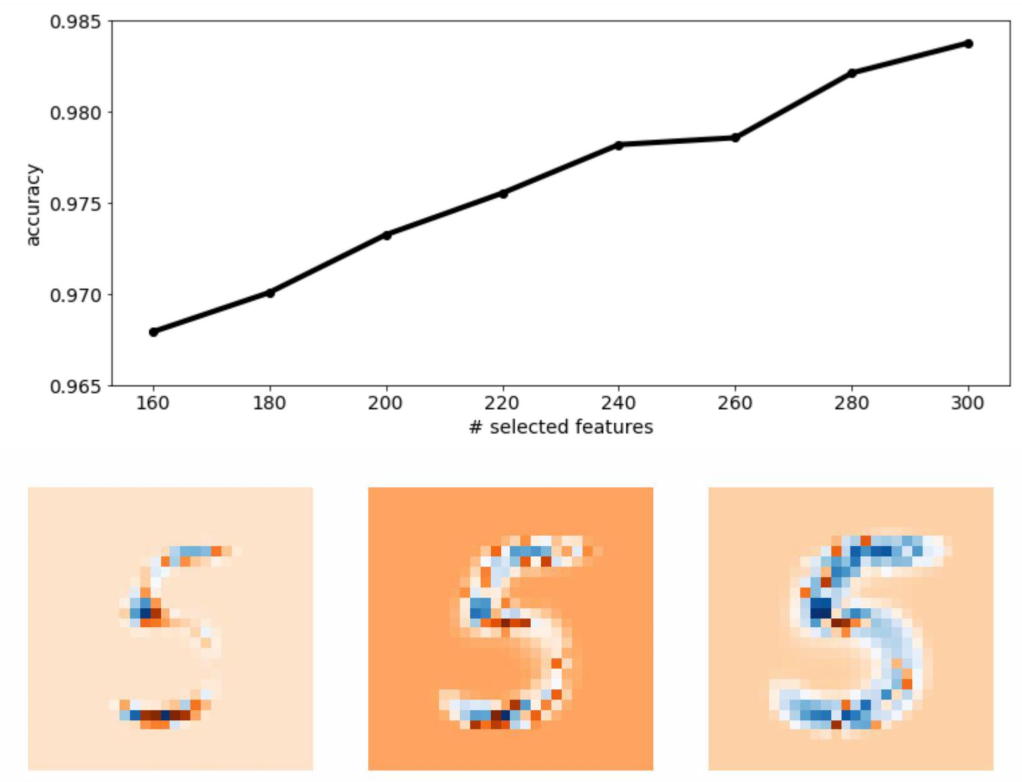
- Consider a data set that consists of the expression levels of various proteins across patients' tissue samples.
- Such measurements are increasingly carried out to assist with disease diagnosis, as biologists measure many proteins with the aim of discriminating between disease classes.
- Yet, it remains expensive to conduct all of the measurements needed to fully characterize proteomic diseases.
- The relationship between the protein expressions and the outcome is sparse and highly nonlinear.




**Feature selection path produced by our method on the MICE Protein Dataset.** The method captures 70% of the signal with about 20% of the features. This allows to narrow down the list of important features, making the conclusions of the prediction task more actionable.

## Background

- **Lasso** (or $\ell_1$-regularized) regression assigns zero weights to the most irrelevant or redundant features, and is widely used in data science.
- The limitation of Lasso, however, is that it only offers solutions to **linear** models.
- We propose a new approach that extends lasso regression and its feature sparsity to feed-forward neural networks.
- The method allows to capture arbitrary nonlinearity in a nonparametric way while simultaneously performing feature selection.



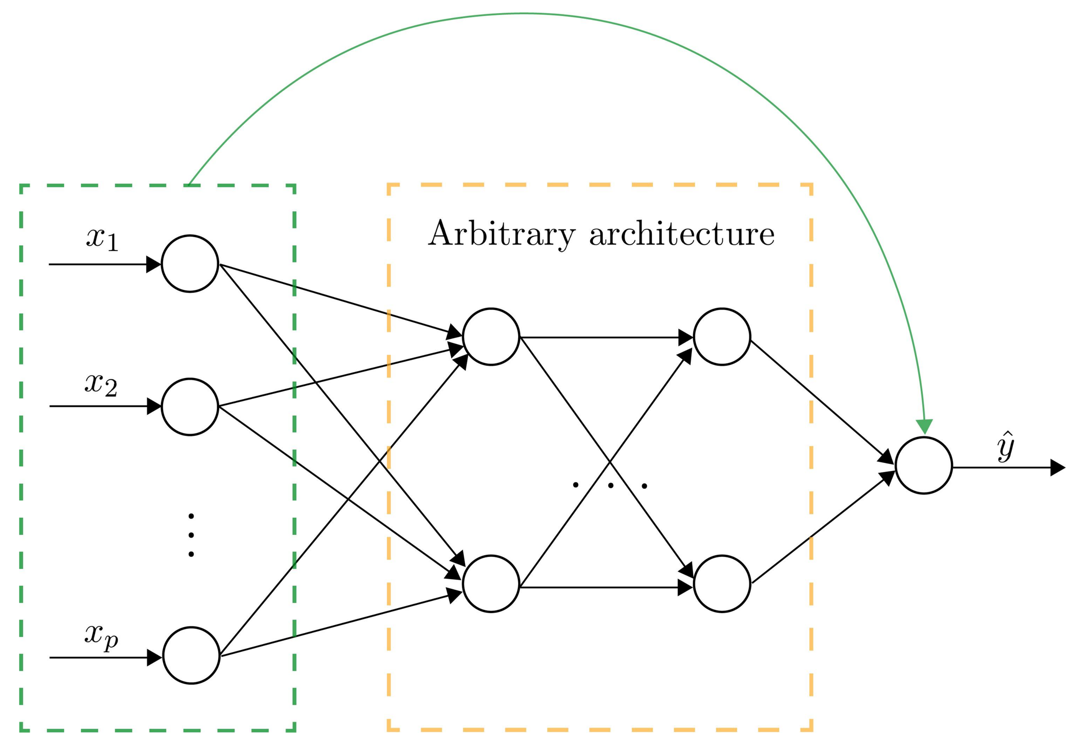

**Demonstrating LassoNet on the MNIST dataset.** Here, we show the results of using LassoNet to simultaneously select informative pixels and classify digits 5 and 6 from the MNIST dataset. *Top:* The classification accuracy by number of selected features. *Bottom:* A sample from the model with 160, 220 and 300 active features (out of the 784 features).

## Formulation of the method

- We consider the class of all fully connected feed-forward residual neural networks, abbreviated as $\mathscr{F} = \left\{f : f(\mathbf{x}) = \theta^T \mathbf{x} + \mathbf{NN}(\mathbf{x}, W)\right\}$.
- **Notation:** $W$ denotes the network parameters, $K$ denotes the size of the first hidden layer, $W^{(0)} \in \mathbb{R}^{d\times K}$ denotes the first hidden layer, $x \in \mathbb{R}^d$ denotes an input data point, and $L(\theta, W) = \frac{1}{n}\sum_{i=1}^{n} \ell(\mathbf{x_i}, y_i; \theta, W)$ denotes the loss on the training data set.
- The LassoNet objective function is defined as

$$\begin{aligned} &\underset{\theta, W}{\text{minimize}} \; L(\theta, W) + \lambda \|\theta\|_1 \\ &\text{subject to } \left\|W_j^{(0)}\right\|_\infty \le M|\theta_j|, \; j = 1, \ldots, d. \end{aligned} \tag{1}$$

- The constraint $|W_{jk}^{(0)}| \le M \cdot |\theta_j|$, $k = 1, \ldots, K$, budgets the total amount of non-linearity involving variable $j$ according to the relative importance of $X_j$ as a linear variable.
- It follows that $W_j = 0$ as soon as $\theta_j = 0$. In other words, variable $j$ is completely inactive from the model, without the need for an explicit penalty on $W$.
- The only hyper-parameters are the hierarchy coefficient, $M$, and the regularization strength, $\lambda$.




**The architecture of LassoNet** consists of a single residual connection, shown in green, and an **arbitrary feed-forward neural network**, shown in **black**. The residual layer and the first hidden layer are optimized jointly using a hierarchical operator.

## Efficient Numerical Optimization

- The objective is optimized using proximal gradient descent.
- The key novelty is a numerically efficient algorithm for the proximal inner-loop.
- The complexity of the algorithm is $O(dK \cdot \log(dK))$, where $d \cdot K$ is the total number of the parameters being updated.
- Thus, the method's complexity is negligible overhead compared to the cost of computing the gradients.

## Competitive Results on Real Data

- We compare LassoNet with several supervised feature selection methods detailed in the paper.
- We measure classification accuracy by passing the selected features to an extremely randomized trees classifier, a variant of random forests that has been used with feature selection methods in prior literature.
- The results are shown here on the ISOLET data set, which is a widely used benchmark for feature selection.
- Overall, we find that our method is the strongest performer in the large majority of cases.