Why do we need Deep Learning?

Matthew Dixon
Department of Applied Mathematics
Illinois Institute of Technology
matthew.dixon@iit.edu

November 4th, 2019

Overview

- Introduction to feedforward neural networks
- Background on approximation and learning theory
- Folding argument for Deep ReLU classifiers[1]

[1] Matus Telgarsky, Representation benefits of deep feedforward network, arXiv:1509.08101, 2015.

Supervised Machine Learning

- Machine learning addresses a fundamental prediction problem: Construct a nonlinear predictor, $\hat{Y}(X)$, of an output, Y, given a high dimensional input matrix $X = (X_1, \ldots, X_P)$ of P variables.
- Machine learning can be simply viewed as the study and construction of an input-output map of the form

$$Y = F(X) \qquad \text{where} \qquad X = (X_1, \ldots, X_P).$$

- The output variable, Y, can be continuous, discrete or mixed.
- For example, in a classification problem, $F : X \to Y$ where $Y \in \{1, \ldots, K\}$ and K is the number of categories.
- We will denote the i^{th} observation of the data $\mathcal{D} := (X, Y)$ as $(\mathbf{x}_i, \mathbf{y}_i)$ - this is a "feature-vector" and response (a.k.a. "label"). Note that lower caps refer to a specific observation in $\mathcal{D}$.

Taxonomy of Most Popular Neural Network Architectures

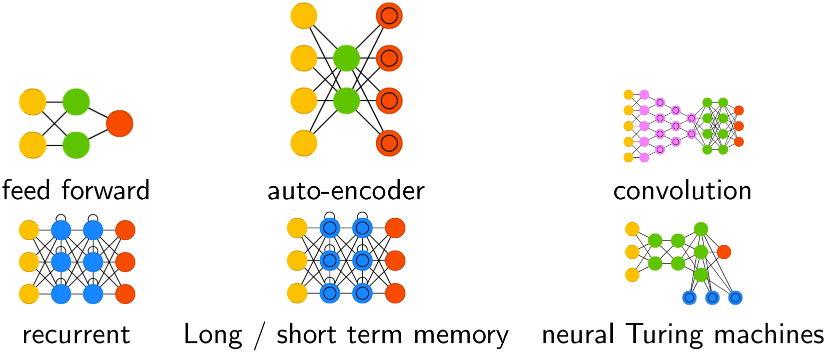


Figure: Most commonly used deep learning architectures for modeling. Source: `http://www.asimovinstitute.org/neural-network-zoo`

FFWD Neural Networks

Neural network model:

$$Y = F_{W,b}(X) + \epsilon$$

where $F_{W,b} : \mathbb{R}^p \to \mathbb{R}^d$ is a deep neural network with L layers

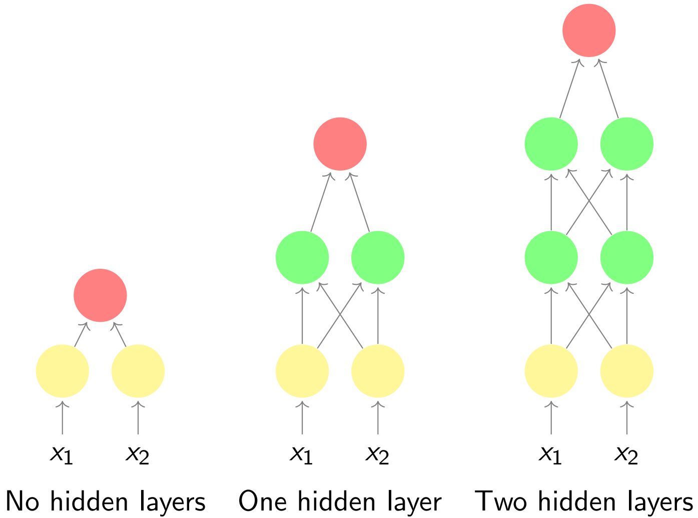
$$\hat{Y}(X) := F_{W,b}(X) = f^{(L)}_{W^{(L)},b^{(L)}} \circ \cdots \circ f^{(1)}_{W^{(1)},b^{(1)}}(X)$$

- $W = (W^{(1)}, \ldots, W^{(L)})$ and $b = (b^{(1)}, \ldots, b^{(L)})$ are weight matrices and bias vectors.
- For any $W^{(i)} \in \mathbf{R}^{m \times n}$, we can write the matrix as n column m-vectors $W^{(i)} = [\mathbf{w}^{(i)}_1, \ldots, \mathbf{w}^{(i)}_n]$.
- Denote each weight as $w^{(\ell)}_{i,j} := \left(W^{(\ell)}\right)_{i,j}$.
- X is a $N \times p$ matrix of observations in $\mathbb{R}^p$.
- No assumptions are made on the distribution of the error, ϵ, other than it is independently distributed.

Deep Neural Networks*

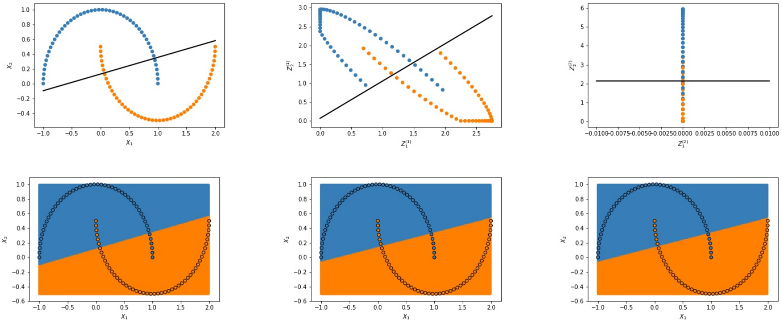
- Let $h : \mathbb{R} \to B \subset \mathbb{R}$ denote a continuous, monotonically increasingly, function.
- A function $f^{(\ell)}_{W^{(\ell)},b^{(\ell)}} : \mathbb{R}^n \to \mathbb{R}^m$, given by $f(v) = W^{(\ell)} h^{(\ell-1)}(v) + b^{(\ell)}$, $W^{(\ell)} \in \mathbb{R}^{m \times n}$ and $b^{(\ell)} \in \mathbb{R}^m$, is a semi-affine function in v, e.g. $f(v) = wtanh(v) + b$.
- $h(\cdot)$ are known activation functions of the output from the previous layer., e.g. $h(x) := \max(x, 0)$ ("ReLU").
- $F_{W,b}(X)$ is a composition of semi-affine functions.
- If all the activation functions are linear, $F_{W,b}$ is just linear regression, regardless of the number of layers L.

Geometric Interpretation of FFWD Neural Networks



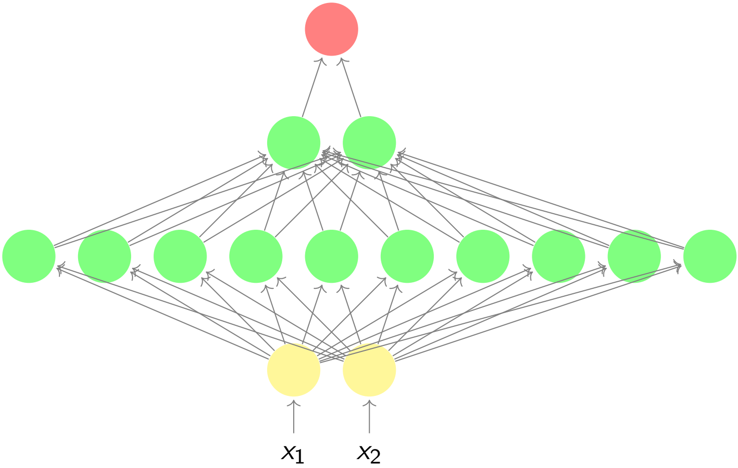
Geometric Interpretation of FFWD Neural Networks

Half-Moon Dataset

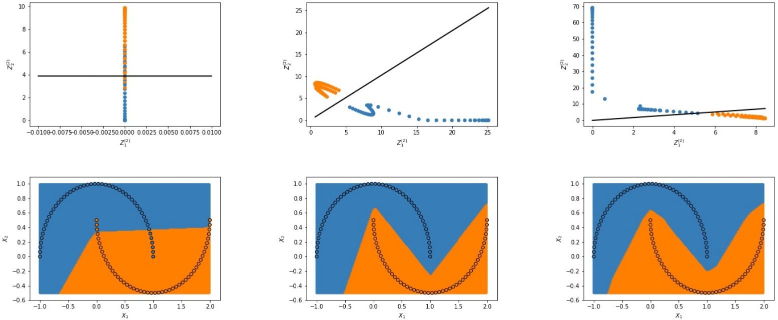


No hidden layers | One hidden layer | Two hidden layers

Why do we need more Neurons?



Geometric Interpretation of Neural Networks



25 hidden units 50 hidden units 75 hidden units

Figure: The number of hidden units is adjusted according to the requirements of the classification problem and can he very high for data sets which are difficult to separate.

Geometric Interpretation of Neural Networks

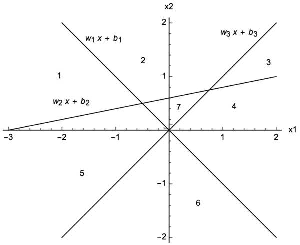


Figure: Hyperplanes defined by three activated neurons in the hidden layer.

Aside: Universal Representation Theorem [1989]

- Let $C^p := \{f : \mathbb{R}^p \to \mathbb{R} \mid f(x) \in C(\mathbb{R})\}$ be the set of continuous functions from $\mathbb{R}^p$ to $\mathbb{R}$.
- Denote $\Sigma^p(h)$ as the class of functions

$$\{F_{W,b} : \mathbb{R}^p \to \mathbb{R} : \; F_{W,b}(x) = W^{(2)} h(W^{(1)} x + b^{(1)}) + b^{(2)}\}.$$

- Consider $\Omega := (0, 1]$ and let $\mathcal{C}_0$ be the collection of all open intervals in $(0, 1]$.
- Then $\sigma(\mathcal{C}_0)$, the σ-algebra generated by $\mathcal{C}_0$, is the Borel σ-algebra, $\mathcal{B}((0, 1])$.
- Let $M^p := \{f : \mathbb{R}^p \to \mathbb{R} \mid f(x) \in \mathcal{B}(\mathbb{R})\}$ denote the set of all Borel measurable functions from $\mathbb{R}^p$ to $\mathbb{R}$.
- Denote the Borel σ-algebra of $\mathbb{R}^p$ as $\mathcal{B}^p$.

Aside: Universal Representation Theorem

Theorem (Hornik, Stinchcombe & White, 1989)

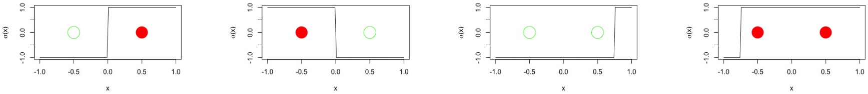
For every monotonically increasing activation function h, every input dimension size p, and every probability measure μ on $(\mathbb{R}^p, \mathcal{B}^p)$, $\Sigma^p(h)$ is uniformly dense on compacta in C^p and ρ_μ-dense in M^p.

URT states that only one hidden layer is needed.... Before we turn to why we need deep classifier networks, let us measure their representational power....

Shattering

- What is the maximum number of points that can be arranged so that $F_{W,b}(X)$ shatters them?
- i.e. for all possible assignments of binary labels to those points, does there exist a W, b such that $F_{W,b}$ makes no errors when classifying that set of data points?
- Every distinct pair of points is separable with the linear threshold perceptron. So every data set of size 2 is shattered by the perceptron.
- However, this linear threshold perceptron is incapable of shattering triplets, for example $X \in \{-0.5, 0, 0.5\}$ and $Y \in \{0, 1, 0\}$.
- In general, the VC dimension of the class of halfspaces in $\mathbb{R}^k$ is $k+1$. For example, a 2-d plane shatters any three points, but can not shatter four points.
- This maximum no. of points is the Vapnik-Chervonenkis (VC) dimension and is one characterization of learnability of a classifier.

Example of Shattering over the Interval $[-1, 1]$



Right point activated | Left point activated | None activated | Both activated

Figure: For the points $\{-0.5, 0.5\}$, there are weights and biases that activates only one of them ($W = 1$, $b = 0$ or $W = -1$, $b = 0$), none of them ($W = 1$, $b = -0.75$) and both of them ($W = 1$, $b = 0.75$).

VC Dimension Example

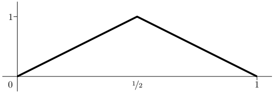
- Determine the VC dimension of the indicator function where $\Omega = [0,1]$

$$\begin{aligned} F(x) &= \{f : \Omega \to \{0,1\},\ f(x) = \mathbb{1}_{x\in[t_1,t_2)},\ \text{or} \\ f(x) &= 1 - \mathbb{1}_{x\in[t_1,t_2)}\ , t_1 < t_2 \in \Omega\}. \end{aligned}$$

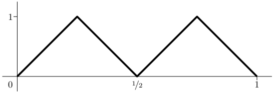
- Suppose there are three points $x_1,\ x_2$ and x_3 and assume $x_1 < x_2 < x_3$ without loss of generality. All possible binary labeling of the points is reachable, therefore we assert that $VC(F) \geq 3$.
- With four points x_1, x_2, x_3 and x_4 (assumed increasing as always), you cannot label x_1 and x_3 with the value 1 and x_2 and x_4 with the value 0 for example. So $VC(F) = 3$.

Single Layer ReLU Networks

$$F_{W,b} = W^{(2)}h(W^{(1)}x + b^{(1)}), \;\; h = \max(x, 0)$$



$2h(x) - 4h(x - \frac{1}{2})$



$4h(x) - 8h(x - \frac{1}{4}) +$
$4h(x - \frac{1}{2}) - 8h(x - \frac{3}{4}).$

Two units

$W^{(2)} = [2, -4]$

$b^{(1)} = [0, -\frac{1}{2}]^T$

Four units

$W^{(2)} = [4, -8, 4, -8]$

$b^{(1)} = [0, -\frac{1}{4}, -\frac{1}{2}, -\frac{3}{4}]^T$

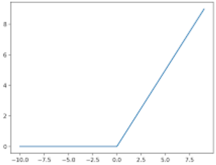
ReLU Networks and Composition

- Consider composing piecewise affine functions instead of adding them.

Definition (t-sawtooth)

$h : \mathbb{R} \to \mathbb{R}$ is t-sawtooth if it is piecewise affine with t pieces, meaning $\mathbb{R}$ is partitioned into t consecutive intervals, and h is affine within each interval.

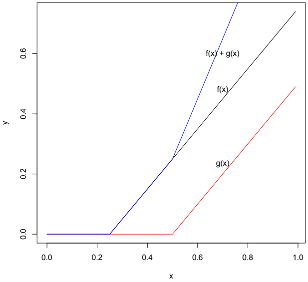
E.g. $ReLU(x)$ is 2-sawtooth,



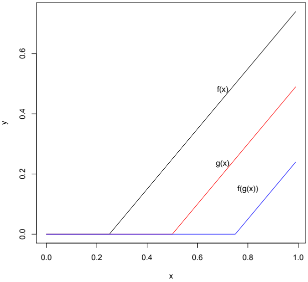
Adding versus composing t-sawtooth functions

Lemma

Let $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$ be respectively k- and l-sawtooth. Then $f + g$ is at most $(k + l)$-sawtooth, and $f \circ g$ is at most kl-sawtooth.



(a) Adding 2-sawtooths

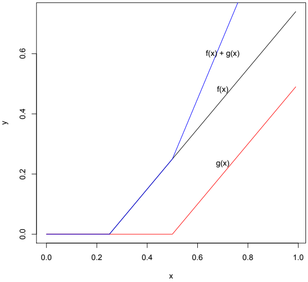


(b) Composing 2-sawtooths

Proof of Adding versus Composing t-sawtooth functions

- Let cl_f denote the partition of $\mathbb{R}$ corresponding to f, and cl_g denote the partition of $\mathbb{R}$ corresponding to g.
- First consider $f+g$: any intervals $U_f \in \mathrm{cl}_f,\ U_g \in \mathrm{cl}_g$.
- $f+g$ has a single slope along $U_f \cap U_g$ and so $f+g$ is $|\mathrm{cl}|$-sawtooth, where cl is the set of all intersections of intervals from cl_f and cl_g, meaning $\mathrm{cl} := \{U_f \cap U_g : U_f \in \mathrm{cl}_f, U_g \in \mathrm{cl}_g\}$.
- By sorting the left endpoints of elements of cl_f and cl_g, it follows that $|\mathrm{cl}| \leq k + l$ (the other intersections are empty).
- Consider the image $f(g(U_g))$ for some interval $U_g \in \mathrm{cl}_g$. g is affine with a single slope along U_g, therefore f is over a single unbroken interval $g(U_g)$.
- Nothing prevents $g(U_g)$ from hitting all the elements of cl_f; since U_g was arbitrary, it holds that $f \circ g$ is at most $(|\mathrm{cl}_f| \cdot |\mathrm{cl}_g|)$-sawtooth.

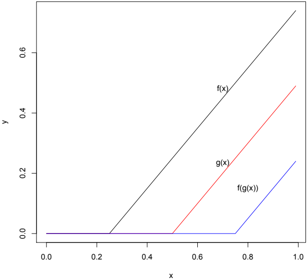
Adding sawtooths



$$f(x) := \max(x - \tfrac{1}{4}, 0), \quad g(x) := \max(x - \tfrac{1}{2}, 0)$$
$$\mathrm{cl}_f = \{[0, \tfrac{1}{4}], (\tfrac{1}{4}, 1]\}, \quad \mathrm{cl}_g = \{[0, \tfrac{1}{2}], (\tfrac{1}{2}, 1]\}.$$

$$\mathrm{cl} = \{[0, \frac{1}{4}] \cap [0, \frac{1}{2}], (\frac{1}{4}, 1] \cap [0, \frac{1}{2}], (\frac{1}{4}, 1] \cap (\frac{1}{2}, 1]\}$$

Composing sawtooths

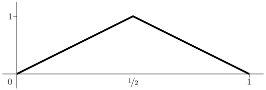


$$f(x) := \max(x - \tfrac{1}{4}, 0), \quad g(x) := \max(x - \tfrac{1}{2}, 0)$$
$$\mathrm{cl}_f = \{[0, \tfrac{1}{4}], (\tfrac{1}{4}, 1]\}, \quad \mathrm{cl}_g = \{[0, \tfrac{1}{2}], (\tfrac{1}{2}, 1]\}.$$

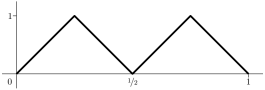
Unfolding

Let us now build on this result by considering the *mirror map* $f_\mathrm{m} : \mathbb{R} \to \mathbb{R}$, which is defined as

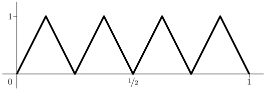
$$f_\mathrm{m}(x) := \begin{cases} 2x & \text{when } 0 \le x \le 1/2, \\ 2(1-x) & \text{when } 1/2 < x \le 1, \\ 0 & \text{otherwise.} \end{cases}$$



f_m



f_m^2



f_m^3.

Unfolding

- Note that f_{m} can be represented by a two layer ReLU activated network with two neurons;
- For instance, $f_{\mathrm{m}}(x) = 2h(x) - 4h(x - 1/2)$.
- Hence f_{m}^k is the composition of k (identical) ReLU sub-networks.

The key observation is that fewer hidden units are needed to shatter a set of points when the network is deep versus shallow....

Shattering Example

Consider for example the sequence of $N := 2^k$ points with alternating labels, referred to as the N-ap.

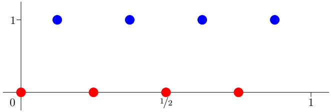


Figure: The N-ap consists of a N uniformly spaced points with alternating labels over the interval $[0, 1 - 2^{-k}]$. That is the points $((x_i, y_i))_{i=1}^{N}$ with $x_i = i2^{-k}$, and $y_i = 0$ when i is even, and otherwise $y_i = 1$. As the x values pass from left to right, the labels change as often as possible and provides the most challenging arrangement for shattering N points.

Classification Error

- Suppose that we have a h activated network with m units per layer and l layers.
- Given a function $f : \mathbb{R}^p \to \mathbb{R}$ let $\tilde{f} : \mathbb{R}^p \to \{0, 1\}$ denote the corresponding classifier $\tilde{f}(x) := \mathbb{1}_{f(x) \geq 1/2}$,
- The data is $((\mathbf{x}_i, y_i))_{i=1}^N$ with $\mathbf{x}_i \in \mathbb{R}^p$ and $y_i \in \{0, 1\}$,
- The error is
$$\mathcal{E}(f) := \frac{1}{N} \sum_i \mathbb{1}_{\tilde{f}(\mathbf{x}_i) \neq y_i}$$
.

Lower Bound on Classification Error

- We have the following lower bound on error of the $t-$sawtooth function for the N-ap:

Lemma

Let $((\mathbf{x}_i, y_i))_{i=1}^N$ be given according to the N-ap. Then every t-sawtooth function $f : \mathbb{R} \to \mathbb{R}$ satisfies $\mathcal{E}(f) \geq (N - 4t)/(3N)$.

- The proof relies on a simple counting argument for the number of crossings of $1/2$.
- If there are m t-saw-tooth functions then by the lemma, the resultant is a piecewise affine function over mt intervals.

Lower Bound on Classification Error

Proof of Lemma.

- Recall the notation $\tilde{f}(x) := [f(x) \geq 1/2]$, whereby $\mathcal{E}(f) := \frac{1}{N}\sum_i [y_i \neq \tilde{f}(x_i)]$.
- Since f is piecewise monotonic with a corresponding partition $\mathbb{R}$ having at most t pieces, then f has at most $2t - 1$ crossings of $1/2$: at most one within each interval of the partition, and at most 1 at the right endpoint of all but the last interval.
- Consequently, $\tilde{f}$ is piecewise *constant*, where the corresponding partition of $\mathbb{R}$ is into at most $2t$ intervals.
- This means N points with alternating labels must land in $2t$ buckets, thus the total number of points landing in buckets with at least three points is at least $N - 4t$.

$\square$

Classification Error

The main theorem now directly follows from the previous lemma.

Theorem

Let positive integer k, number of layers l, and number of nodes per layer m be given. Given a t-sawtooth $h : \mathbb{R} \to \mathbb{R}$ and $N = 2^k$ points as specified by the N-ap, then

$$\min_{W,b} \mathcal{E}(f) \geq \frac{N - 4(tm)^l}{3N}.$$

Summary

- The theorem states that ReLU function composition is more efficient than function addition
- From this result one can say, for example, that on the N-ap one needs $m = 2^{k-3}$ many units to perfecting classifying with a ReLU activated shallow network versus only $m = 2^{\frac{k-(l+2)}{l}}$ units per layer for a $l \geq 2$ deep ReLU network.