Introduction to Deep Neural Networks

Numerical Methods for Deep Learning

Why Deep Networks?

- Universal approximation theorem of NN suggests that we can approximate **any** function by two layers.
- But - The width of the layer can be very large $\mathcal{O}(n \cdot n_f)$
- Deeper architectures can lead to more efficient descriptions of the problem.
- No real proof but lots of practical experience.

Deep Neural Networks

How deep is deep?
We will answer this question later ...

Until recently, the standard architecture was

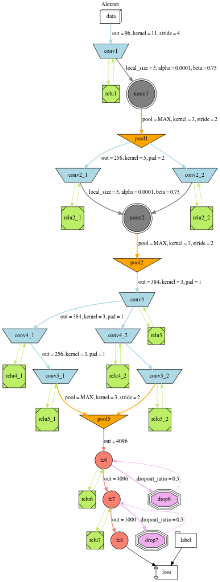
$$\begin{aligned}\mathbf{Y}_1 &= \sigma(\mathbf{K}_0\mathbf{Y}_0 + b_0)\\ \vdots &= \vdots\\ \mathbf{Y}_N &= \sigma(\mathbf{K}_{N-1}\mathbf{Y}_{N-1} + b_{N-1})\end{aligned}$$

And use $\mathbf{Y}_N$ to classify. This leads to the optimization problem

$$\min_{\mathbf{K}_{0,\ldots,N-1},\mathbf{b}_{0,\ldots,N-1},\mathbf{W}} E\left(\mathbf{W}\mathbf{Y}_N(\mathbf{K}_1,\ldots,\mathbf{K}_{N-1},b_1,\ldots,b_{N-1}),\mathbf{C}^{\mathrm{obs}}\right)$$

Example: The Alexnet [?] for Image Classification

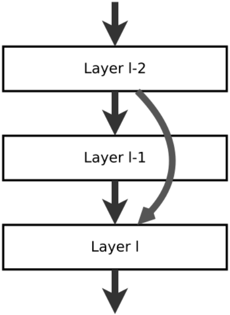
- Complex architectures
- trained on multiple GPUs
- $\approx$ 60 million weights



Deep Neural Networks in Practice

(Some) challenges:

- Computational costs (architecture have millions or billions of parameters)
- difficult to design
- difficult to train (exploding/vanishing gradients)
- unpredictable performance



In 2015, He et al. [**? ?**] came with a new architecture that solves many of the problems

Simplified Residual Neural Network

Residual Network

$$\begin{aligned}\mathbf{Y}_1 &= \mathbf{Y}_0 + \sigma(\mathbf{K}_0 \mathbf{Y}_0 + b_0) \\ \vdots &= \vdots \\ \mathbf{Y}_N &= \mathbf{Y}_{N-1} + \sigma(\mathbf{K}_{N-1} \mathbf{Y}_{N-1} + b_{N-1})\end{aligned}$$

And use $\mathbf{Y}_N$ to classify. This leads to the optimization problem

$$\min_{\mathbf{K}_{0,\ldots,N-1}, \mathbf{b}_{0,\ldots,N-1}, \mathbf{W}} E\left(\mathbf{W}\mathbf{Y}_N(\mathbf{K}_1, \ldots, \mathbf{K}_{N-1}, b_1, \ldots, b_{N-1}), \mathbf{C}^{\mathrm{obs}}\right)$$

Leads to smoother objective function [**?**].

Stability of Deep Residual Networks

Why are ResNets more stable?
A small change

$$
\begin{aligned}
\mathbf{Y}_1 &= \mathbf{Y}_0 + h\sigma(\mathbf{K}_0\mathbf{Y}_0 + b_0) \\
\vdots &= \vdots \\
\mathbf{Y}_N &= \mathbf{Y}_{N-1} + h\sigma(\mathbf{K}_{N-1}\mathbf{Y}_{N-1} + b_{N-1})
\end{aligned}
$$

This is nothing but a forward Euler discretization of the Ordinary Differential Equation (ODE)

$$
\dot{\mathbf{Y}}(t) = \sigma(\mathbf{K}(t)\mathbf{Y}(t) + b(t)), \qquad \mathbf{Y}(0) = \mathbf{Y}_0
$$

We can understand the behavior by learning the dynamics of nonlinear ODEs [**? ?**].

Crash Course on ODEs

Given the ODE

$$\dot{\mathbf{y}} = f(t, \mathbf{y})$$

Assumptions:

1. f differentiable with Jacobian

$$\mathbf{J}(t, \mathbf{y}) = \left(\frac{\partial f}{\partial \mathbf{y}}\right)^\top$$

2. $\mathbf{J}$ changes sufficiently slowly in time

Then (see also [**?** **?** **?**])

- If $Re(\mathrm{eig}(\mathbf{J})) > 0$ $\rightarrow$ Unstable
- If $Re(\mathrm{eig}(\mathbf{J})) < 0$ $\rightarrow$ Stable (converge to a stationary point)
- If $Re(\mathrm{eig}(\mathbf{J})) = 0$ $\rightarrow$ Stable, energy bounded

Stability of Residual Network

Assume forward propagation of single example $\mathbf{y}_0$

$$\dot{\mathbf{y}}(t) = \sigma(\mathbf{K}(t)\mathbf{y}(t) + b(t)), \qquad \mathbf{y}(0) = \mathbf{y}_0$$

The Jacobian is

$$\mathbf{J}(t) = \mathrm{diag}\left(\sigma'(\mathbf{K}(t)\mathbf{y}(t) + b(t))\right)\mathbf{K}(t)$$

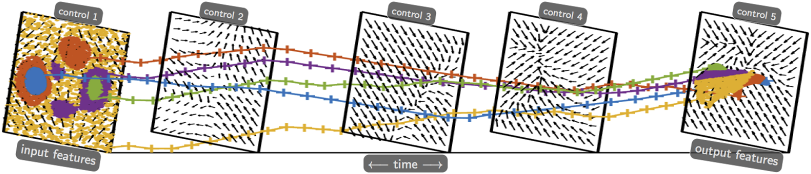
Here, $\sigma'(x) \geq 0$ for tanh, ReLU, ...

Hence, problem is stable when

1. $\mathbf{J}$ changes slowly in time
2. $Re(\mathrm{eig}\mathbf{K}(t)) \leq 0$ for every t

Remember that we learn $\mathbf{K}$ $\rightsquigarrow$ ensure stability by regularization/constraints!

Residual Network as a Path Planning Problem



Forward propagation in residual network (continuous)

$$\dot{\mathbf{Y}}(t) = \sigma(\mathbf{K}(t)\mathbf{Y}(t) + b(t)), \qquad \mathbf{Y}(0) = \mathbf{Y}_0$$

The goal is to plan a path (via $\mathbf{K}$ and b) such that the initial data can be linearly separated

Question: What is a layer, what is depth?

Stability: Continuous vs. Discrete

Assume $\mathbf{K}$ is chosen so that the (continuous) forward propagation is stable

$$\dot{\mathbf{Y}}(t) = \sigma(\mathbf{K}\mathbf{Y}(t) + b(t)), \qquad \mathbf{Y}(0) = \mathbf{Y}_0$$

And assume we use the forward Euler method to discretize

$$\mathbf{Y}_{j+1} = \mathbf{Y}_j + h\sigma(\mathbf{K}_j\mathbf{Y}_j + b_j)$$

Is the network stable?

Not always ...

Stability: A Simple Example

Look at the simplest possible forward propagation

$$\dot{\mathbf{Y}}(t) = \lambda \mathbf{Y}(t)$$

And assume we use the forward Euler to discretize

$$\mathbf{Y}_{j+1} = \mathbf{Y}_j + h\lambda \mathbf{Y}_j = (1 + h\lambda)\mathbf{Y}_j$$

Then the method is stable only if

$$|1 + h\lambda| \leq 1$$

Not every network is stable! Time step size depends on our Jacobian

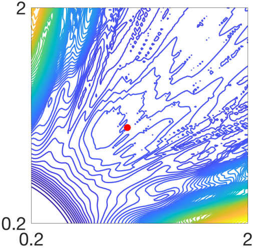
Why you should care about stability - 1

$$\min_{\theta} \frac{1}{2}\|\mathbf{Y}_N(\theta) - \mathbf{C}\|_F^2 \qquad \mathbf{Y}_{j+1}(\theta) = \mathbf{Y}_j(\theta) + \frac{10}{N}\tanh\left(\mathbf{K}\mathbf{Y}_j(\theta)\right)$$

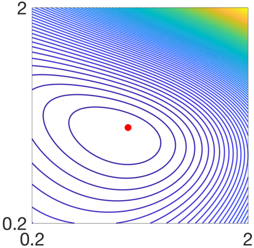
where $\mathbf{C} = \mathbf{Y}_{200}(1,1)$, $\mathbf{Y}_0 \sim \mathcal{N}(0,1)$, and

$$\mathbf{K}(\theta) = \begin{pmatrix} -\theta_1 - \theta_2 & \theta_1 & \theta_2 \\ \theta_2 & -\theta_1 - \theta_2 & \theta_1 \\ \theta_1 & \theta_2 & -\theta_1 - \theta_2 \end{pmatrix}$$

objective, $N = 5$

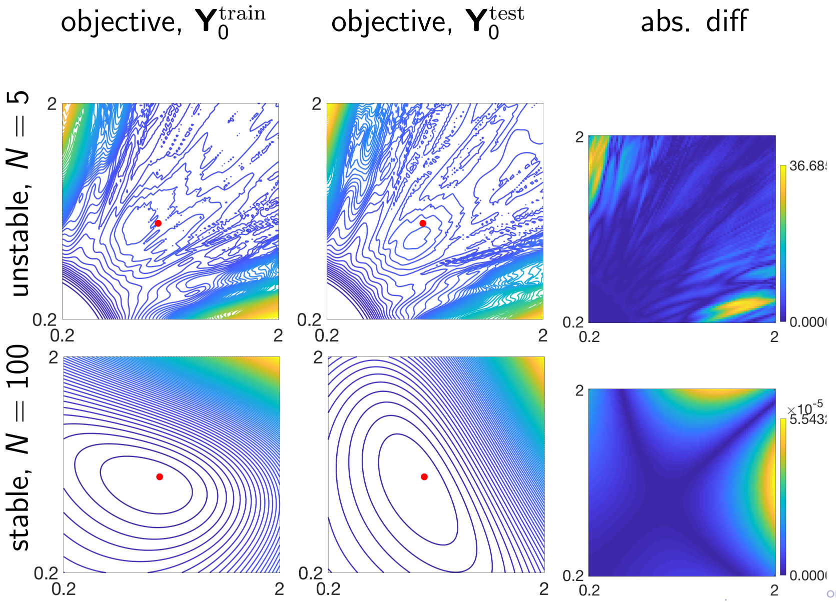


objective, $N = 100$



Next: Compare different inputs $\sim$ generalization

Why you should care about stability - 2



Stability: A Non-Trivial Example

Consider the antisymmetric kernel model

$$\mathbf{K}(t) = \mathbf{K}(t) - \mathbf{K}(t)^\top$$

Here, $Re(\mathrm{eig}(\mathbf{J})(t)) = 0$ for all θ.

Assume we use the forward Euler to discretize

$$\mathbf{Y}_{j+1} = \mathbf{Y}_j + h\sigma((\mathbf{K}_j - \mathbf{K}_j^\top)\mathbf{Y}_j + b_j)$$

Tricky question: How to pick h to ensure stability?
Answer: Impossible since eigenvalues of Jacobian are imaginary. Need other method than forward Euler.

References