A generalized neural tangent kernel for surrogate gradient learning

Spotlight Talk, NeurIPS 2024

Luke Eilers [1,2]   Raoul-Martin Memmesheimer [3]   Sven Goedeke [3,4]

[1]Department of Physiology, University of Bern

[2]Institute for Applied Mathematics, University of Bonn

[3]Institute of Genetics, University of Bonn

[4]Bernstein Center Freiburg, University of Freiburg

Motivation

- Neural network training methods require a useful gradient of the network function

- Activation functions with discontinuities are common in theoretical neuroscience
  $\hookrightarrow$ Gradient methods cannot be used!

- Consider the infinite-width limit, where the neural tangent kernel (NTK) describes the dynamics of gradient descent
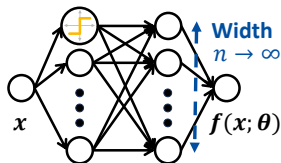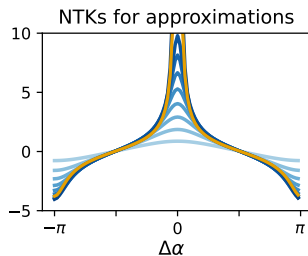
No NTK for step activation functions

- Approximate the step activation function with smooth functions

Activation functions σ
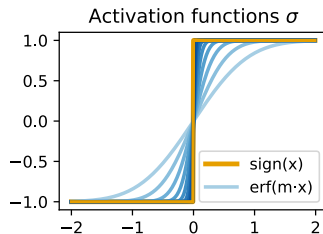
- The NTK diverges as the approximation gets closer ↪ No NTK for step activation functions

NTKs for approximations

The SG-NTK for surrogate gradient learning

- *Surrogate gradient learning*: successful learning of networks without useful gradient
- We derive the **surrogate gradient NTK (SG-NTK)**
- We can define an SG-NTK also for step activation functions

SG-NTKs for approximations
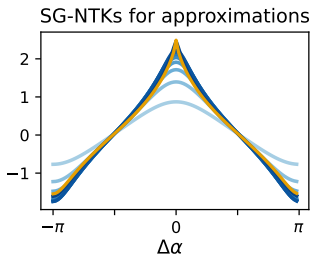
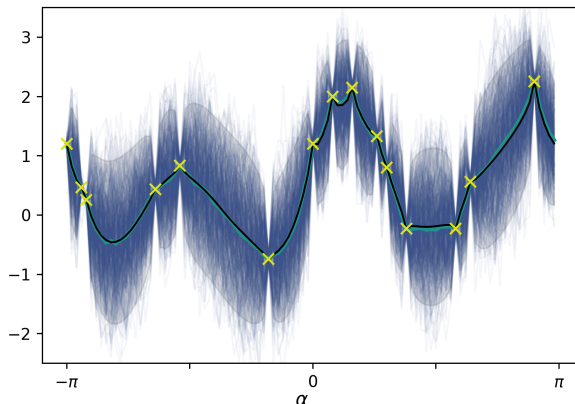- Can the SG-NTK actually describe networks trained with surrogate gradient learning and step activation function?

Figure 1: 500 networks (blue) with step function as activation function and hidden layer widths $n = 500$ trained with SGL for $t = 3\mathrm{e}4$ time steps plotted together with their mean (cyan), the SG-NTK-GP's mean (black), and confidence band (grey). Training points are indicated with crosses.