NTK - The Kernel Bible 2
Jacobian Analysis

Axel Bjarkar Sigurjónsson
Tristan Þórðarson†

10. febrúar 2024

1 Introduction

The last Kernel bible failed to account for all ANNs where the output is not a scalar. In the Kernel Bible 2, we will observe how the NTK kernel is computed if our network function can be written as:

$$f : \mathbb{R}^n \to \mathbb{R}^m \tag{1}$$

2 Setup

If equation 1 holds true, then f_θ can me modeled as such:

$$f(x) = \begin{bmatrix} f_1(x_1, \dots, x_n) \\ \vdots \\ f_m(x_1, \dots, x_n) \end{bmatrix} \tag{2}$$

Where each $f_i \quad i = 1, 2, \dots, n$ is a function described by the models parameters. All parameters (individual weights and biases) are stored in the θ vector:

$$\theta = \begin{bmatrix} W^{[1]}_{1,1} \\ \vdots \\ B^{[1]}_{1,1} \\ \vdots \\ W^{[p]}_{1,1} \\ \vdots \\ B^{[p]}_{[1]} \\ \vdots \\ B^{[p]}_{[m]} \end{bmatrix} \tag{3}$$

Where $p - 1$ is number of hidden layers

2.1 Premise

Previously, the model returned a scalar so our Kernel was constructed from individuals gradients:

$$K_{i,j} = \nabla f_\theta(x_i)^T \nabla f_\theta(x_j)$$

But if our model can be written to the format of equation 2, our entry for the Kernel becomes:

$$K_{i,j} = \begin{bmatrix} \sum_k \frac{\partial f_1(x_i)}{\partial\theta_k}\frac{\partial f_1(x_j)}{\partial\theta_k} & \cdots & \sum_k \frac{\partial f_1(x_i)}{\partial\theta_k}\frac{\partial f_m(x_j)}{\partial\theta_k} \\ \vdots & \ddots & \vdots \\ \sum_k \frac{\partial f_m(x_i)}{\partial\theta_k}\frac{\partial f_1(x_j)}{\partial\theta_k} & \cdots & \sum_k \frac{\partial f_m(x_i)}{\partial\theta_k}\frac{\partial f_m(x_j)}{\partial\theta_k} \end{bmatrix} \tag{4}$$

Where k is the number of parameters in θ.

So K is defined to a be Block matrix with each $K_{i,j} \in \mathbb{R}^{m\times m}$.
Let N represent the number of datapoints. Then $K \in \mathbb{R}^{N(m\times m)}$

2.2 Jacobian Approach

Let $Df_\theta(x)$ represent the Jacobian matrix of our model f with respect to all parameters θ, evaluated at datapoint x:

$$Df_\theta(x) = \underbrace{\begin{bmatrix} \frac{\partial f_1(x)}{\partial\theta_1} & \cdots & \frac{\partial f_m(x)}{\partial\theta_1} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_1(x)}{\partial\theta_k} & \cdots & \frac{\partial f_m(x)}{\partial\theta_k} \end{bmatrix}}_{k\times m} \tag{5}$$

Then each entry of K can greatly simplified:

$$K_{i,j} = [Df_\theta(x_i)]^T Df_\theta(x_j) \tag{6}$$

And therefore:

$$K = \begin{bmatrix} [Df_\theta(x_1)]^T Df_\theta(x_1) & \cdots & Df_\theta(x_1)]^T Df_\theta(x_N) \\ \vdots & \ddots & \vdots \\ [Df_\theta(x_N)]^T Df_\theta(x_1) & \cdots & Df_\theta(x_N)]^T Df_\theta(x_N) \end{bmatrix} \tag{7}$$