Introduction

In the study of neural networks, particularly in the analysis of initialization, understanding the behavior of the kernel is crucial. The kernel $K^{(\ell)}_{\alpha\beta}$ represents the covariance of preactivations at layer ℓ. This analysis is vital for ensuring that the network's parameters are initialized in such a way that the network can train efficiently and effectively. This paper focuses on two types of susceptibilities, parallel and perpendicular, which play critical roles in understanding the stability and dynamics of neural networks at initialization. We will explore the definitions, critical behaviors, and the universality classes associated with these susceptibilities.

1 Parallel Susceptibility

The kernel $K^{(\ell)}_{\alpha_1\alpha_2}$ is defined in equation (4.116) as follows:

$$\mathbb{E}\left[z^{(\ell)}_{i_1;\alpha_1} z^{(\ell)}_{i_2;\alpha_2}\right] = \delta_{i_1 i_2} G^{(\ell)}_{\alpha_1\alpha_2} = \delta_{i_1 i_2}\left[K^{(\ell)}_{\alpha_1\alpha_2} + O\left(\frac{1}{n}\right)\right],$$

where $z^{(\ell)}_{i;\alpha}$ is the α-th component of the i-th hidden unit in the ℓ-th layer, and $G^{(\ell)}_{\alpha_1\alpha_2}$ is the Gram matrix of the hidden units in the ℓ-th layer. The kernel $K^{(\ell)}_{\alpha_1\alpha_2}$ is defined as the limit of the Gram matrix $G^{(\ell)}_{\alpha_1\alpha_2}$ as the number of hidden units n goes to infinity. The kernel $K^{(\ell)}_{\alpha_1\alpha_2}$ is a function of the weights and biases of the network and is therefore defined by the following equality:

$$K^{(\ell)}_{\alpha_1\alpha_2} + O\left(\frac{1}{n}\right) = G^{(\ell)}_{\alpha_1\alpha_2}$$

To study how this kernel evolves with depth, we consider its recursion relation:

$$K^{(\ell+1)}_{\alpha\beta} = C_b + C_W \left\langle \sigma(z_\alpha)\sigma(z_\beta)\right\rangle_{K^{(\ell)}},$$

where:

- C_b is the variance of the biases,
- C_W is the variance of the weights,
- σ is the activation function,
- $\langle\cdot\rangle_{K^{(\ell)}}$ denotes the expectation with respect to the distribution of z_α and z_β, which are Gaussian random variables with covariance $K^{(\ell)}$.

To understand the stability and critical behavior of this recursion, we introduce the concept of *parallel susceptibility*. The parallel susceptibility $\chi_\parallel$ measures the sensitivity of the kernel to perturbations around its fixed point.

First, we define the helper function $g(K)$:

$$g(K) = \frac{1}{\sqrt{2\pi K}} \int_{-\infty}^{\infty} dz\, e^{-\frac{z^2}{2K}} \sigma(z)^2.$$

Using this, the kernel recursion can be written as:

$$K_{\alpha\beta}^{(\ell+1)} = C_b + C_W g(K_{\alpha\beta}^{(\ell)}).$$

Fixed Point Analysis

To find the fixed point K_0^* of this recursion, we solve:

$$K_0^* = C_b + C_W g(K_0^*).$$

Linear Stability Analysis

Next, we perform a linear stability analysis around the fixed point. We consider a small perturbation $\delta K^{(\ell)}$ around the fixed point K_0^*:

$$K^{(\ell)} = K_0^* + \delta K^{(\ell)}.$$

Substituting this into the recursion relation, we get:

$$K_0^* + \delta K^{(\ell+1)} = C_b + C_W g(K_0^* + \delta K^{(\ell)}).$$

Expanding $g(K)$ around K_0^* to first order, we have:

$$g(K_0^* + \delta K^{(\ell)}) \approx g(K_0^*) + g'(K_0^*)\delta K^{(\ell)},$$

where $g'(K)$ is the derivative of $g(K)$ with respect to K.

Thus, the recursion relation for the perturbation becomes:

$$K_0^* + \delta K^{(\ell+1)} \approx C_b + C_W \left[g(K_0^*) + g'(K_0^*)\delta K^{(\ell)} \right].$$

Using the fixed point condition $K_0^* = C_b + C_W g(K_0^*)$, we simplify this to:

$$\delta K^{(\ell+1)} \approx C_W g'(K_0^*)\delta K^{(\ell)}.$$

This shows that the perturbation evolves according to:

$$\delta K^{(\ell+1)} = \chi_{\parallel} \delta K^{(\ell)},$$

where $\chi_{\parallel} = C_W g'(K_0^*)$ is the parallel susceptibility.

Expression for $g'(K)$

The derivative $g'(K)$ is given by:

$$g'(K) = \frac{d}{dK}\left[\frac{1}{\sqrt{2\pi K}}\int_{-\infty}^{\infty} dz\, e^{-\frac{z^2}{2K}}\sigma(z)^2\right].$$

Using the Leibniz rule for differentiation under the integral sign and some algebra, we get:

$$g'(K) = \frac{1}{2K}\left[\frac{1}{\sqrt{2\pi K}}\int_{-\infty}^{\infty} dz\, e^{-\frac{z^2}{2K}}\sigma(z)^2\left(\frac{z^2}{K}-1\right)\right].$$

Thus, the parallel susceptibility becomes:

$$\chi_{\parallel} = C_W\frac{1}{2K_0^*}\left[\frac{1}{\sqrt{2\pi K_0^*}}\int_{-\infty}^{\infty} dz\, e^{-\frac{z^2}{2K_0^*}}\sigma(z)^2\left(\frac{z^2}{K_0^*}-1\right)\right].$$

Criticality Condition

To maintain a stable kernel, the perturbations should neither grow unbounded nor decay to zero. This leads to the criticality condition:

$$\chi_{\parallel}(K_0^*) = 1,$$

which ensures that $\delta K^{(\ell)}$ does not diverge or vanish, maintaining a stable kernel value.

In summary, the parallel susceptibility $\chi_{\parallel}$ quantifies the sensitivity of the kernel recursion to small perturbations around the fixed point. It is a crucial parameter in determining the stability and critical behavior of neural networks at initialization.

2 Perpendicular Susceptibility

The perpendicular susceptibility $\chi_{\perp}$ measures the sensitivity of the kernel to perturbations that are perpendicular to the original inputs. It arises in the context of analyzing how differences between inputs propagate through the network.

Consider a perturbation $\delta K_{[2]}^{(\ell)}$ that represents the difference between two distinct inputs. The recursion for this perturbation can be written as:

$$\delta K_{[2]}^{(\ell+1)} = \chi_{\perp}\delta K_{[2]}^{(\ell)},$$

where $\chi_{\perp}$ is the perpendicular susceptibility.

We define $\chi_{\perp}$ as:

$$\chi_{\perp}(K) \equiv C_W\left\langle \sigma'(z)^2\right\rangle_K,$$

where $\sigma'(z)$ is the derivative of the activation function.

This expectation can be expressed as:

$$\left\langle \sigma'(z)^2 \right\rangle_K = \frac{1}{\sqrt{2\pi K}} \int_{-\infty}^{\infty} dz\, e^{-\frac{z^2}{2K}} \sigma'(z)^2.$$

Therefore, the perpendicular susceptibility is given by:

$$\chi_\perp(K) = C_W \frac{1}{\sqrt{2\pi K}} \int_{-\infty}^{\infty} dz\, e^{-\frac{z^2}{2K}} \sigma'(z)^2.$$

3 Comparison of Susceptibilities

Both susceptibilities, parallel and perpendicular, play crucial roles in the stability analysis of neural networks at initialization.

- **Parallel Susceptibility** $\chi_\parallel$: Measures the sensitivity to perturbations parallel to the input and is given by:

$$\chi_\parallel = C_W \frac{1}{2K_0^*} \left[\frac{1}{\sqrt{2\pi K_0^*}} \int_{-\infty}^{\infty} dz\, e^{-\frac{z^2}{2K_0^*}} \sigma(z)^2 \left(\frac{z^2}{K_0^*} - 1 \right) \right].$$

  It ensures that the perturbations neither grow unbounded nor decay to zero by satisfying the criticality condition $\chi_\parallel(K_0^*) = 1$.

- **Perpendicular Susceptibility** $\chi_\perp$: Measures the sensitivity to perturbations perpendicular to the input and is given by:

$$\chi_\perp(K) = C_W \frac{1}{\sqrt{2\pi K}} \int_{-\infty}^{\infty} dz\, e^{-\frac{z^2}{2K}} \sigma'(z)^2.$$

  It controls the growth or decay of differences between inputs as they propagate through the network.

In summary, both susceptibilities are essential for understanding the dynamics and stability of neural networks at initialization. The parallel susceptibility ensures that the overall scale of the kernel remains stable, while the perpendicular susceptibility governs the propagation of differences between distinct inputs.

4 Universality Classes

Scale-Invariant Universality Class

For scale-invariant activation functions, such as the ReLU activation, both the parallel and perpendicular susceptibilities are equal and independent of the kernel. This property greatly simplifies the criticality analysis. The susceptibilities are given by:

$$\chi_\parallel(K) = \chi_\perp(K) = \chi,$$

where χ is a constant.

This equality means that the critical initialization hyperparameters (C_b, C_W) can be found by solving the simpler set of equations:

$$K_{00}^{(\ell+1)} = C_b + C_W g(K_{00}^{(\ell)}),$$

$$g(K) = \langle \sigma(z)\sigma(z) \rangle_K ,$$

$$\chi_{\parallel}(K) = C_W g'(K) = 1,$$

$$\chi_{\perp}(K) = C_W \langle \sigma'(z)\sigma'(z) \rangle_K = 1.$$

Thus, for scale-invariant activation functions, the critical initialization hyperparameters are:

$$(C_b, C_W)_{\text{critical}} = \left(0, \frac{1}{\sigma_1^2}\right),$$

where σ_1 is the first derivative of the activation function at zero.

Non-Scale-Invariant Universality Class

For non-scale-invariant activation functions, such as tanh or sin, the parallel and perpendicular susceptibilities are not equal and must be analyzed independently. The susceptibilities are given by:

$$\chi_{\parallel}(K) = \frac{C_W}{2K^2} \left\langle \sigma'(z)\sigma'(z)(z^2 - K) \right\rangle_K ,$$

$$\chi_{\perp}(K) = C_W \langle \sigma'(z)\sigma'(z) \rangle_K .$$

For these functions, the critical initialization hyperparameters (C_b, C_W) are found by solving the more complex set of equations:

$$K_{00}^{(\ell+1)} = C_b + C_W g(K_{00}^{(\ell)}),$$

$$g(K) = \langle \sigma(z)\sigma(z) \rangle_K ,$$

$$\chi_{\parallel}(K) = 1,$$

$$\chi_{\perp}(K) = 1.$$

In this case, the critical initialization hyperparameters are:

$$(C_b, C_W)_{\text{critical}} = \left(0, \frac{1}{\sigma_1^2}\right),$$

with the additional condition that $\sigma_0 = 0$ and $\sigma_1 \neq 0$. This condition ensures that the activation function has a non-trivial fixed point at $K_{00}^* = 0$.

Comparison of Universality Classes

1. **Scale-Invariant Universality Class**:
   - Both susceptibilities are equal and independent of the kernel:
     $$\chi_{\parallel}(K) = \chi_{\perp}(K) = \chi.$$
   - Critical initialization hyperparameters are easier to determine.
2. **Non-Scale-Invariant Universality Class**:
   - Susceptibilities differ and must be analyzed separately:
     $$\chi_{\parallel}(K) = \frac{C_W}{2K^2} \left\langle \sigma'(z)\sigma'(z)(z^2 - K) \right\rangle_K ,$$
     $$\chi_{\perp}(K) = C_W \left\langle \sigma'(z)\sigma'(z) \right\rangle_K .$$
   - Critical initialization hyperparameters are determined by more complex conditions:
     $$(C_b, C_W)_{\text{critical}} = \left(0, \frac{1}{\sigma_1^2}\right),$$
     with $\sigma_0 = 0$ and $\sigma_1 \neq 0$.

In conclusion, the universality class for scale-invariant activation functions is characterized by the equality of parallel and perpendicular susceptibilities, simplifying the analysis. For non-scale-invariant functions, the susceptibilities are different, requiring a more detailed examination to determine the critical initialization hyperparameters.