Neural Network Characteristic Functions

Evan Dogariu*

May 3, 2024

Abstract

{evan: write something}

1 Background

For notation, let $\mathbb{R}_+ := [0, \infty)$.

1.1 Neural Network

We will proceed first for general MLPs, and then specifically focus on those under NTK parameterization. The goal will be to determine layerwise dynamics that we can meaningfully analyze in the infinite-depth limit. For now, suppose that there are no biases and the network has scalar output.

Definition 1 (Neural Network)**.** *A neural network with depth $L \geq 2$, input width n_0, hidden width n, and elementwise activation function $\sigma : \mathbb{R} \to \mathbb{R}$ is given by*

$$z_{\ell+1}^{(\alpha)} := \sqrt{\frac{c}{n}} W_\ell \sigma(z_\ell^{(\alpha)}) \quad (\forall \ell \in \{1, \dots, L\}) \tag{$\dagger$}$$

where $W_\ell \in \begin{cases} \mathbb{R}^{n \times n} & \ell \in \{1, \dots, L-1\} \\ W_\ell \in \mathbb{R}^{n \times n_0} & \ell = 0 \\ W_\ell \in \mathbb{R}^{1 \times n} & \ell = L \end{cases}$ *are the weights, $z_\ell^{(\alpha)}$ denotes the preactivations at layer ℓ when passing in input $x_\alpha \in \mathbb{R}^{n_0}$, and so $z_{L+1}^{(\alpha)} \in \mathbb{R}$ is the model's output. We initialize with $(W_\ell)_{i,j} \sim \mathcal{N}(0, 1)$ i.i.d. (with the factor of $\sqrt{c/n}$ this amounts to the He initialization where c is determined by σ).*

Note that, for example, we can apply the NTK parameterization by defining $\sigma := \frac{1}{\sqrt{n}}\phi$ for an activation function ϕ that is independent of n. For activation shaping with an initial smooth activation function $\phi : \mathbb{R} \to \mathbb{R}$, we would define

$$\sigma_a(z) := a\sqrt{n} \cdot \phi\left(\frac{z}{a\sqrt{n}}\right) \quad (\forall z \in \mathbb{R}^n)$$

*Princeton University

for some constant $a > 0$ (see Definition 3.6 in [1]). For now, we proceed for general σ.

1.2 Characteristic Functions

Let $(\Omega, \mathcal{M}, \mathbb{P})$ be a probability space.

Definition 2 (Characteristic Function)**.** *Let $X : \Omega \to \mathbb{R}^n$ a random variable. We define the* characteristic function *of X to be $\widehat{f}_X : \mathbb{R}^n \to \mathbb{C}$ given by*

$$\widehat{f}(t) := \mathbb{E}\left[e^{i\langle t, X\rangle}\right] = \int_{x \in \mathbb{R}^n} e^{i\langle t, x\rangle} f_X(x),$$

where the last equality holds if X is absolutely continuous and so its density f_X exists.

Note the relation to the Fourier transform. We have the following properties:

- $\widehat{f}_X$ exists and is uniformly continuous for all random variables X (even singular ones).
- For all $t \in \mathbb{R}^n$, $|\widehat{f}_X(t)| \leq 1$ and $\widehat{f}_X(0) = 1$.
- $\widehat{f}_X \in L^2(\mathbb{R}^n)$ if (and only if) the density f_X exists and is square-integrable.
- If the distribution of X is rotationally-symmetric, then $\widehat{f}_X$ is always a real-valued and even function.
- $\widehat{f}_X(at)\widehat{f}_Y(bt) = \widehat{f}_{(aX+bY)}(t)$; this comes from the relationship between convolution and the Fourier transform.

2 Single Input

2.1 Dynamics

For this subsection, we fix a single input $x \equiv x_\alpha$ and drop the α labels for notation. Note that for each layer $\ell \in \{1, \ldots, L+1\}$, the value of $(z_\ell)_j$ (the j^{th} coordinate of the preactivation) on initialization is a real-valued random variable (furthermore, by rotational symmetry of the normal distribution, it is the same for all coordinates j). Denote by $\varphi_\ell : \mathbb{R} \to \mathbb{C}$ the characteristic function of the preactivations at layer ℓ for a single fixed input: in other words,

$$\varphi_\ell(t) := \mathbb{E}\left[e^{it(z_\ell)_j}\right] \quad (\forall t \in \mathbb{R})$$

where this has the same value for all j. For a smooth activation we expect the distribution of z_ℓ to be a.c., and so $\varphi_\ell \in L^2(\mathbb{R})$ for each layer ℓ. Furthermore, by rotational invariance of the normal distribution, φ_ℓ is always real-valued.

Now, let's relate $\varphi_{\ell+1}$ to φ_ℓ using the dynamics (†). In particular, for any coordinate $j \in \{1, \ldots, n\}$ we have

$$(z_{\ell+1})_j = \sqrt{\frac{c}{n}} \sum_{k=1}^{n} (W_\ell)_{j,k} \cdot \sigma((z_\ell)_k)$$

Letting $Y_{\ell,j,k} : \Omega \to \mathbb{R}$ denote the independent (since weights are drawn i.i.d.) real-valued random variables $(W_\ell)_{j,k} \cdot \sigma((z_\ell)_k)$, we get the following relationship between the characteristic functions:

$$\varphi_{\ell+1}(t) = \prod_{k=1}^{n} \widehat{f}_{Y_{\ell,j,k}}\left(t\sqrt{\frac{c}{n}}\right)$$

Furthermore,

$$\begin{aligned}
\widehat{f}_{Y_{\ell,j,k}}(s) &= \mathbb{E}\left[e^{is(W_\ell)_{j,k}\cdot\sigma((z_\ell)_k)}\right] \\
&= \int_{w\in\mathbb{R}} \int_{z\in\mathbb{R}} e^{isw\sigma(z)} \frac{1}{2\pi} \int_{u\in\mathbb{R}} e^{-iuz} \varphi_\ell(u) du dz \frac{e^{-w^2/2}}{\sqrt{2\pi}} dw
\end{aligned}$$

where we used that $(W_\ell)_{j,k}$ is a standard normal and $(z_\ell)_k$ is the random variable with density given by the inverse Fourier transform of φ_ℓ (such a density exists because the preactivations are absolutely continuous). Note that the above expression does not depend on k. So, we can plug this into what we had earlier to see

$$(2\pi)^{3/2} \varphi_{\ell+1}(t)^{1/n} = \int_{\mathbb{R}^3} e^{-w^2/2 + it\sigma(z)\sqrt{c/n}w - iuz} \varphi_\ell(u) du dz dw$$

Using the Gaussian integral relation $\int_{\mathbb{R}} e^{-(ax^2+bx+c)} = \sqrt{\frac{\pi}{a}} e^{-c+b^2/4a}$, we can integrate out w to see

$$\begin{aligned}
(2\pi)\varphi_{\ell+1}(t)^{1/n} &= \int_{\mathbb{R}^2} e^{-iuz - ct^2(\sigma(z)^2)/2n} \varphi_\ell(u) du dz \\
&= \int_{u\in\mathbb{R}} \varphi_\ell(u) \int_{z\in\mathbb{R}} e^{-iuz - ct^2\sigma^2(z)/2n} dz du
\end{aligned}$$

Noting that $\varphi_\ell(u) = \varphi_\ell(-u)$ by rotational symmetry, we arrive at

$$\varphi_{\ell+1}(t)^{1/n} = \frac{1}{\pi} \int_0^\infty \varphi_\ell(u) \int_{z\in\mathbb{R}} \cos(uz) e^{-ct^2\sigma^2(z)/2n} dz du$$

Using the relation $\widehat{f}_X(nz) = \widehat{f}_{(nX)}(z) = \widehat{f}_{X+\ldots+X}(z) = \left(\widehat{f}_X(z)\right)^n$, this can be written

$$\varphi_{\ell+1}(t) = \frac{1}{\pi} \int_0^\infty \varphi_\ell(u) \int_{z\in\mathbb{R}} \cos(uz) e^{-cnt^2\sigma^2(z)/2} dz du$$

Consider the map $K_n : \mathbb{R}_+ \times \mathbb{R}_+ \to \mathbb{R}$ given by

$$K_n(t,u) = \frac{1}{\pi} \int_{z \in \mathbb{R}} \cos(uz) e^{-cnt^2 \sigma^2(z)/2} dz,$$

which is explicitly calculable given only the activation function σ. Let $\mathcal{H} := L^2(\mathbb{R}_+)$ denote the Hilbert space of square-integrable functions from $\mathbb{R}_+ \to \mathbb{R}$, and form the integral operator $T_{K,n} : \mathcal{H} \to \mathcal{H}$ given by

$$T_{K,n}(f) := \int_0^\infty f(u) K_n(\cdot, u) du$$

$T_{K,n}$ is a bounded linear operator, and since K_n is a function that is in $L^2(\mathbb{R}_+ \times \mathbb{R}_+)$ it even holds that $T_{K,n}$ is a Hilbert-Schmidt integral operator, which is therefore compact. Each φ_ℓ is an element of $\mathcal{H}$, and so we have the dynamics

$$\varphi_{\ell+1} = T_{K,n}(\varphi_\ell)$$

with initialization $\phi_1(t) = e^{-t^2 c \|x\|^2/2}$ (where $x \in \mathbb{R}^{n_0}$ was the input data point).

2.2 Spectral Properties of $T_{K,n}$

Lemma 1. *The spectral theorem applies to $T_{K,n} \in \mathcal{B}(\mathcal{H})$.*

Proof. We will show that $T_{K,n}$ is a normal operator. So, let $f \in \mathcal{H} \equiv L^2(\mathbb{R}_+)$. We want to verify that $T_{K,n} T_{K,n}^*(f) = T_{K,n}^* T_{K,n}(f)$. We know

$$\begin{aligned} T_{K,n} T_{K,n}^*(f) &= \int_0^\infty K_n(\cdot, u) \int_0^\infty K_n(v,u) f(v) dv du \\ &= \int_0^\infty f(v) \int_0^\infty K_n(\cdot, u) K_n(v,u) du dv \end{aligned}$$

and

$$\begin{aligned} T_{K,n}^* T_{K,n}(f) &= \int_0^\infty K_n(u, \cdot) \int_0^\infty K_n(u,v) f(v) dv du \\ &= \int_0^\infty f(v) \int_0^\infty K_n(u, \cdot) K_n(u,v) du dv \end{aligned}$$

So, it suffices to show that for Lebesgue a.e. $t, v \in [0, \infty)$,

$$\int_0^\infty K_n(t,u) K_n(v,u) du = \int_0^\infty K_n(u,t) K_n(u,v) du$$

{evan: do this} □

2.3 NTK Parameterization

In the NTK parameterization, for a given initial activation function $\sigma : \mathbb{R} \to \mathbb{R}$ we apply update (†) with activation function $\sigma/\sqrt{n}$ instead. Plugging this into our machinery, we see that

$$K(t,u) \equiv K_n(t,u) = \frac{1}{\sqrt{\pi}} \int_{z \in \mathbb{R}} \cos(uz) e^{-ct^2 \sigma^2(z)/2} dz$$

is independent of n! So, letting T_K denote the resulting compact linear operator on $\mathcal{H}$, we have the infinite-dimensional linear dynamical system on $\mathcal{H}$ given by

$$\varphi_{\ell+1} = T_K \varphi_\ell$$

with

$$\varphi_1(\cdot) = e^{-c\|x\|^2 (\cdot)^2/2}$$

This is really all we need to describe the distribution of the preactivation of each individual neuron, at least at large depths.

References

[1] neural covariance SDE paper