Neural Network Characteristic Functions

Evan Dogariu*

May 3, 2024

Abstract

{evan: write something. the goal is to compute the correlation statistics of infinite-depth networks in a new way}

Contents

1 THINGS TO DO (in order of importance) 1

2 Background 2
2.1 Neural Network . . . 2
2.2 Characteristic Functions . . . 2

3 Dynamics 3
3.1 Single Input . . . 3
3.2 Two Inputs . . . 5
3.3 Many Inputs . . . 6

4 Properties of K_n 8
4.1 Spectral Properties . . . 8
4.2 Forming an ODE - WIP . . . 9
4.3 Examples . . . 10
4.3.1 Deep Linear Networks . . . 10

1 THINGS TO DO (in order of importance)

1. Prove that $T_{K,n}$ is a normal operator?
2. Write out an ODE that the eigenfunctions η_j or γ_j satisfy, and figure out bounds on λ_j.

*Princeton University

2 Background

2.1 Neural Network

We will proceed first for general MLPs, and then specifically focus on those under NTK parameterization. The goal will be to determine layer-wise 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*

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

where $W_\ell \in \begin{cases} \mathbb{R}^{n \times n} & \ell \in \{1, \ldots, L-1\} \\ W_\ell \in \mathbb{R}^{n \times n_0} & \ell = 0 \\ W_\ell \in \mathbb{R}^{1 \times n} & \ell = L \end{cases}$ *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

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

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

2.2 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.

3 Dynamics

First, we will derive the dynamics of the characteristic functions for a single input, from which we will generalize to multiple inputs and arrive at Proposition 1.

3.1 Single Input

For this subsection, we fix a single input $x \equiv x_\alpha$ and drop the α labels for notation. Let $\mathbb{R}_+$ denote $[0, \infty)$. 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,

$$\varphi_\ell(t) := \mathbb{E}\left[e^{itz_{\ell,j}}\right] \quad (\forall t \in \mathbb{R})$$

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,

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

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

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

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

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

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

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

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

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

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. Form the integral operator $T_{K,n} : L^2(\mathbb{R}_+)\to L^2(\mathbb{R}_+)$ 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 $L^2(\mathbb{R}_+)$, and so we have the dynamics

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

with initialization $\varphi_1(t) = e^{-t^2c\|x\|^2/2}$ (where $x\in\mathbb{R}^{n_0}$ was the input data point).

3.2 Two Inputs

Next, we are interested in the joint distributions of the preactivations of a neuron for *two different inputs*. So, fix two different inputs $x_\alpha, x_\beta \in \mathbb{R}^{n_0}$ and for each $\ell \in \{1, \ldots, L+1\}$ consider the $\mathbb{R}^2$-valued random variable $(z_\ell^{(\alpha)}, z_\ell^{(\beta)})$. Denote by $\psi_\ell^{(\alpha,\beta)} : \mathbb{R}^2 \to \mathbb{C}$ the characteristic function of this random variable: in other words,

$$\psi_\ell^{(\alpha,\beta)}(t) := \mathbb{E}\left[e^{i\left\langle t, \left(z_{\ell,j}^{(\alpha)}, z_{\ell,j}^{(\beta)}\right)\right\rangle_{\mathbb{R}^2}}\right] \quad (\forall t \in \mathbb{R}^2)$$

where as before this has the same value for all j due to rotational symmetry[1]. For a smooth activation we expect the distribution of $(z_\ell^\alpha, z_\ell^\beta)$ to be a.c. w.r.t. $\mathbb{R}^2$, and so $\psi_\ell \in L^2(\mathbb{R}^2)$ for each layer ℓ. Furthermore, by rotational invariance of the normal distribution, ψ_ℓ is always real-valued.

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

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

Letting $Z_{\ell,j,k} : \Omega \to \mathbb{R}^2$ denote the independent[2] $\mathbb{R}^2$-valued random variables $\left((W_\ell)_{j,k} \cdot \sigma(z_{\ell,k}^{(\alpha)}),\ (W_\ell)_{j,k} \cdot \sigma(z_{\ell,k}^{(\beta)})\right)$, we get the following relationship between the characteristic functions:

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

Furthermore,

$$\begin{aligned} \widehat{f}_{Z_{\ell,j,k}}(s) &= \mathbb{E}\left[e^{is_1 (W_\ell)_{j,k} \cdot \sigma(z_{\ell,k}^{(\alpha)}) + is_2 (W_\ell)_{j,k} \cdot \sigma(z_{\ell,k}^{(\beta)})}\right] \\ &= \int_{w \in \mathbb{R}} \int_{z \in \mathbb{R}^2} e^{iw\langle s, \sigma(z)\rangle} \frac{1}{(2\pi)^2} \int_{u \in \mathbb{R}^2} e^{-i\langle u, z\rangle} \psi_\ell(u) du dz \frac{e^{-w^2/2}}{\sqrt{2\pi}} dw \end{aligned}$$

where we used that $(W_\ell)_{j,k}$ is a standard normal and $\left(z_{\ell,k}^{(\alpha)},\ z_{\ell,k}^{(\beta)}\right)$ is the random variable with density given by the inverse Fourier transform of ψ_ℓ (such a density exists because the joint 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

$$(2\pi)^{5/2} \psi_{\ell+1}(t)^{1/n} = \int_{\mathbb{R}^5} e^{-w^2/2 + i\sqrt{c/n}w\langle t, \sigma(z)\rangle - i\langle u, z\rangle} \psi_\ell(u) du dz dw$$

[1] Note also that the random variables $z_{\ell,i}^\alpha$ and $z_{\ell,j}^\beta$ are always independent if $i \neq j$. So, we only study the case $i = j$.

[2] Here, I mean that Z_{ℓ_1,j_1,k_1} and Z_{ℓ_2,j_2,k_2} are independent if any of $\ell_1 \neq \ell_2, j_1 \neq j_2, k_1 \neq k_2$ occur. As before, these are independent since weights are drawn i.i.d.

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

$$\begin{aligned}(2\pi)^2 \psi_{\ell+1}(t)^{1/n} &= \int_{\mathbb{R}^4} e^{-i\langle u,z\rangle - c\langle t,\sigma(z)\rangle^2/2n} \psi_\ell(u) du dz \\ &= \int_{u\in\mathbb{R}^2} \psi_\ell(u) \int_{z\in\mathbb{R}} e^{-i\langle u,z\rangle - c\langle t,\sigma(z)\rangle^2/2n} dz du\end{aligned}$$

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

$$\psi_{\ell+1}(t) = \frac{1}{(2\pi)^2} \int_{u\in\mathbb{R}^2} \psi_\ell(u) \int_{z\in\mathbb{R}^2} e^{-i\langle u,z\rangle - cn\langle t,\sigma(z)\rangle^2/2} dz du$$

Consider the map $K_n : \mathbb{R}^2 \times \mathbb{R}^2 \to \mathbb{R}$ given by

$$K_n(t,u) = \frac{1}{(2\pi)^2} \int_{z\in\mathbb{R}^2} e^{-i\langle u,z\rangle - cn\langle t,\sigma(z)\rangle^2/2} dz,$$

which is explicitly calculable given only the activation function σ. Form the integral operator $T_{K,n} : L^2(\mathbb{R}^2) \to L^2(\mathbb{R}^2)$ given by

$$T_{K,n}(f) := \int_{u\in\mathbb{R}^2} f(u) K_n(\cdot, u) du$$

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

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

with initialization $\psi_1(t) = e^{-t_1^2 c\|x_\alpha\|^2/2} e^{-t_2^2 c\|x_\beta\|^2/2}$ (where $x_\alpha, x_\beta \in \mathbb{R}^{n_0}$ were the input data points).

3.3 Many Inputs

From the previous subsection, it is obvious how to generalize this. We restate all assumptions up until this point as well.

Definition 3. *Consider a neural network (as per Definition 1) with width n, depth L, input dimension n_0, and activation function σ. Let $\{x_\alpha\}_{\alpha=1}^m \subseteq \mathbb{R}^{n_0}$ be a finite* dataset. *The random $\mathbb{R}^m$-valued variable $Z_{\ell,j} := \left(z_{\ell,j}^{(1)}, \ldots, z_{\ell,j}^{(m)}\right)$ has the same distribution for all j if ℓ is held fixed, and $Z_{\ell,j}$ has characteristic function*

$$\psi_\ell : \mathbb{R}^m \to \mathbb{C} \quad \textit{sending } t \mapsto \mathbb{E}_{z\sim Z_{\ell,j}}\left[e^{i\langle t,z\rangle_{\mathbb{R}^m}}\right]$$

*We call ψ_ℓ the m-*point state *of the layer-wise dynamical system. By taking the Fourier transform of ψ_ℓ (if $\psi_\ell \in L^2$), one has the density of the joint distribution of $Z_{\ell,j}$, from which any m-wise statistic may be computed.*

Proposition 1 (Dynamics)**.** *Let $(\psi_\ell)_{\ell=1}^{L+1}$ be as given in Definition 3. If $\sigma \in C^1(\mathbb{R})$, then $\psi_\ell \in L^2(\mathbb{R}^m \to \mathbb{C})$ and the layer-wise dynamics of $(\psi_\ell)_\ell$ are given by the following discrete-time, time-invariant linear dynamical system on $L^2(\mathbb{R}^m \to \mathbb{C})$: for $\ell \in \{1, \dots, L\}$,*

$$\psi_{\ell+1} = T_{K,n}\psi_\ell := \int_{u\in\mathbb{R}^m} K_n(\cdot, u)\psi_\ell(u)du$$

with initial value

$$\psi_1(t) = e^{-c\sum_{k=1}^m t_k^2\|x_k\|^2/2} \quad (\forall t \in \mathbb{R}^m)$$

In the above, $K_n \in L^2(\mathbb{R}^m \times \mathbb{R}^m \to \mathbb{R})$ is a square-integrable (non-symmetric) kernel function given by

$$K_n(t, u) := \frac{1}{(2\pi)^m}\int_{z\in\mathbb{R}^m} e^{-i\langle u,z\rangle - cn\langle t,\sigma(z)\rangle^2/2}dz$$

and $T_{K,n}$ is the resulting compact Hilbert-Schmidt integral operator.

Corollary 1 (m-Point Correlations)**.** *Let $\mathcal{F}$ denote the Fourier transform on $L^2(\mathbb{R}^m \to \mathbb{C})$ If $\sigma \in C^1$, then for all $\ell \in \{1, \dots, L+1\}$ and all $j \in \{1, \dots, n\}$,*

$$\mathbb{E}\left[\prod_{k=1}^m z_{\ell,j}^{(k)}\right] = \int_{u\in\mathbb{R}^m}\left(\prod_{k=1}^m u_k\right)\left(\mathcal{F}^* T_{K,n}^{\ell-1}\psi_1\right)(u)du$$

Proof. The random variable $\prod_{k=1}^m z_{\ell,j}^{(k)}$ is absolutely continuous {evan: why?}. So, a Fourier inversion yields the joint density of the random variables $(z_{\ell,j}^{(1)}, \dots, z_{\ell,j}^{(m)})$, with which we compute the expectation in the ordinary way. □

In order to get the correlation statistics, it suffices to figure out enough about $T_{K,n}^{\ell-1}\psi_1$ in order to compute the above integral. Since we are interested in large ℓ this will, of course, require understanding the spectral properties of $T_{K,n}$ – this is performed in Section 4.1. We will primarily be interested in how things change with n, and especially joint scalings of $n, L \to \infty$.

Remark 1. In the NTK parameterization, for a given initial activation function $\sigma : \mathbb{R} \to \mathbb{R}$ one applies update (†) with activation function $\sigma/\sqrt{n}$ instead. Plugging this into our machinery, we see that

$$K(t, u) \equiv K_n(t, u) = \frac{1}{(2\pi)^m}\int_{z\in\mathbb{R}^m} e^{-i\langle u,z\rangle - c\langle t,\sigma(z)\rangle^2/2}dz$$

is independent of n. This sheds light on why for fixed depth, the infinite-width limit is stable in the NTK regime. However, if one were to consider infinite-depth (with *any* widths, finite or infinite) under the NTK parameterization, the characteristic functions would converge to the delta at 0 (because of the e^{-t^2} effect of K), which is equivalent to the distributions of preactivations converging toward being uniform. This is an uninteresting infinite-depth limit.

4 Properties of K_n

4.1 Spectral Properties

In this subsection, we start with some analytic properties of the kernel, after which we study spectral properties of the operator $T_{K,n}$.

Lemma 1 (Derivatives of K_n)**.** *The map $K_n : \mathbb{R}^m \times \mathbb{R}^m \to \mathbb{C}$ is smooth in each coordinate, with first derivatives*

$$(\nabla_t K_n)(t,u) = \frac{cn}{(2\pi)^m} \int_{z\in\mathbb{R}^m} \sigma(z)\, \langle t, \sigma(z)\rangle\, e^{-i\langle u,z\rangle - cn\langle t,\sigma(z)\rangle^2/2} dz$$

and

$$(\nabla_u K_n)(t,u) = \frac{-i}{(2\pi)^m} \int_{z\in\mathbb{R}^m} z e^{-i\langle u,z\rangle - cn\langle t,\sigma(z)\rangle^2/2} dz$$

Proof. This follows from iterating dominated convergence and the chain rule. Nothing too special, and higher derivatives can be calculated similarly. □

The rub is that although $T_{K,n}$ is not self-adjoint {evan: though the jury is still out on if it is normal}, it is compact and so we know that any nonzero elements of $\sigma(T_{K,n})$ are eigenvalues.

Proposition 2 (Spectral Properties of $T_{K,n}$ and $T^*_{K,n}$)**.** *Let $T_{K,n} \in \mathcal{B}(L^2(\mathbb{R}^m \to \mathbb{C}))$ be as given in Proposition 1. Then, there is an orthonormal set $(\eta_j)_j \subseteq L^2(\mathbb{R}^m \to \mathbb{C})$ of eigenfunctions of $T_{K,n}$ with eigenvalues $\sigma(T_{K,n}) = (\lambda_j)_j \subseteq \mathbb{C}$ satisfying $\lambda_j \to 0$. Furthermore, each η_j is a.e. infinitely-differentiable with*

$$\nabla \eta_j(t) = \lambda_j \int_{z\in\mathbb{R}^m} \eta_j(z) \nabla_t K_n(t,z) dz$$

Similarly, there is an orthonormal set $(\gamma_j)_j \subseteq L^2(\mathbb{R}^m \to \mathbb{C})$ of a.e. smooth eigenfunctions of $T^*_{K,n}$ with eigenvalues $\sigma(T^*_{K,n}) = (\overline{\lambda_j})_j \subseteq \mathbb{C}$ satisfying

$$\nabla \gamma_j(z) = \overline{\lambda_j} \int_{t\in\mathbb{R}^m} \gamma_j(t) \nabla_z K_n(t,z) dt$$

If $T_{K,n}$ is normal (commutes with its adjoint), then $\{\eta_j\}_j$, $\{\gamma_j\}_j$ even form orthonormal bases of $L^2(\mathbb{R}^m \to \mathbb{C})$.

Proof. As $T_{K,n}$ is compact, the spectral theory of compact operators (see the statement here) tells us that there is an orthonormal set $(\eta_j)_j \subseteq L^2(\mathbb{R}^m \to \mathbb{C})$ of eigenfunctions of $T_{K,n}$ with eigenvalues $(\lambda_j)_j \subseteq \mathbb{C}$ satisfying $\lambda_j \to 0$. The eigenvector condition reads

$$\eta_j(t) = \lambda_j \int_{u\in\mathbb{R}^m} \eta_j(u) K_n(t,u) du$$

Since K_n is exponentially-decaying in t (and so Lipschitz in t for fixed u), we see that η_j is a.e. differentiable by Rademacher's theorem. Taking the derivative and applying dominated convergence,

$$\nabla \eta_j(t) = \lambda_j \int_{u \in \mathbb{R}^m} \eta_j(u) \nabla_t K_n(t,u) du$$

By Lemma 1 and similar logic to the above, $\nabla \eta_j$ is (locally-)Lipschitz as a function of t, and so we can take another derivative. This can be repeated for the higher derivatives, and so we expect that η_j is infinitely-differentiable. Similar properties hold for the adjoint, though in this instance we are dealing with iterated derivatives of $\cos(\cdot)$ (which are still Lipschitz, though decay slower than $e^{-(\cdot)^2}$). The last statement comes from the spectral theory for normal operators. □

4.2 Forming an ODE - WIP

Combining (0) - (2), we get the ODE

$$\eta_j''(t) + \frac{1}{t}\eta_j'(t) = t^2 \lambda_j \int_{z \in \mathbb{R}} g(z)^2 A_j(t,z) dz$$

We note that $A_j(t,z) = \frac{1}{\pi} e^{-g(z)t^2/2} \langle \eta_j, \cos(z(\cdot)) \rangle_{\mathcal{H}}$, and so (0) gives

$$\eta_j(t) = \frac{\lambda_j}{\pi} \int_{z \in \mathbb{R}} \langle \eta_j, \cos(z \cdot) \rangle \, e^{-g(z)t^2/2} dz$$

Defining $W_j(w) := \langle \eta_j, \cos(w(\cdot)) \rangle_{\mathcal{H}}$ and integrating the above expression,

$$\begin{aligned} W_j(z) &= \frac{\lambda_j}{\pi} \int_0^\infty \cos(zt) \int_{w \in \mathbb{R}} W_j(w) e^{-g(w)t^2/2} dw dt \\ &= \frac{\lambda_j}{\pi} \int_{w \in \mathbb{R}} W_j(w) \int_0^\infty \cos(zt) e^{-g(w)t^2/2} dt dw \end{aligned}$$

Using the Gaussian integral identity $\int_0^\infty \cos(at) e^{-bt^2} dt = \sqrt{\frac{\pi}{4b}} e^{-a^2/4b}$, we get that

$$W_j(z) = \frac{\lambda_j}{\sqrt{2\pi c n}} \int_{w \in \mathbb{R}} \frac{W_j(w)}{\sigma(w)} e^{-z^2/2cn\sigma^2(w)} dw$$

and therefore that

$$A_j(t,z) = \frac{\lambda_j}{\pi} e^{-g(z)t^2/2} \int_{w \in \mathbb{R}} \frac{W_j(w)}{\sqrt{2\pi g(w)}} e^{-z^2/2g(w)} dw$$

Plugging this into (0),

$$\begin{aligned} \eta_j(t) &= \frac{\lambda_j^2}{\pi} \int_{z \in \mathbb{R}} \int_{w \in \mathbb{R}} e^{-g(z)t^2/2} \frac{W_j(w)}{\sqrt{2\pi g(w)}} e^{-z^2/2g(w)} dw dz \\ &= \frac{\lambda_j^2}{\pi} \int_{w \in \mathbb{R}} W_j(w) \int_{z \in \mathbb{R}} e^{-g(z)t^2/2} \frac{e^{-z^2/2g(w)}}{\sqrt{2\pi g(w)}} dz dw \end{aligned}$$

The inner z integral is $\mathbb{E}_{z\sim\mathcal{N}(0,g(w))}\left[e^{-g(z)t^2/2}\right] = \mathbb{E}_{z\sim\mathcal{N}(0,1)}\left[e^{-g\left(z/\sqrt{g(w)}\right)t^2/2}\right]$

{evan: **keep going?**}

4.3 Examples

The math above is a bit unenlightening. Let's focus on common examples to really get a sense for how things look.

4.3.1 Deep Linear Networks

Consider the setting where $\sigma(t) = t$, i.e. linear activations. In this case, the kernel has the simpler form

$$\begin{aligned} K_n(t,u) &= \frac{1}{(2\pi)^m}\int_{z\in\mathbb{R}^m} e^{-i\langle u,z\rangle - cn\langle t,z\rangle^2/2}dz \\ &\{evan: \textbf{KEEP GOIN with generalizing to } m \textbf{ points}\} \\ &= \sqrt{\frac{2}{cnt^2}}e^{-u^2/2cnt^2} \end{aligned}$$

Note that for any $f \in L^2([0,\infty)) \equiv \mathcal{H}$ and a.e. $s \in [0,\infty)$,

$$(T_{K,n}f)\left(\sqrt{\frac{1}{cns^2}}\right) = s\sqrt{2}\int_0^\infty f(u)e^{-u^2s^2/2}du$$

Rescaling (and inverting) the domain is always a (diagonal and invertible) linear operation on $\mathcal{H}$, and the integral operator $f \mapsto \int_0^\infty fe^{-(\cdot)^2s^2/2}$ is self-adjoint, compact, and easily diagonalizable. Furthermore, multiplying the function by s can be viewed as applying the unbounded position operator X on $\mathcal{H}$. $T_{K,n}$ is therefore self-adjoint, compact (by the two-sided-$*$-ideal property), and easily diagonalizable, from which we can describe the evolution of φ_ℓ as $\ell \to \infty$.

References

[1] neural covariance SDE paper