6

Stability and neural tangent kernel

6.1 Stability in the non-mean field regime

So far, we only considered the mean field regime, where the state of every layer can be exactly determined by conditional expectation of the previous layer. However, when width d is finite, because the sample and population means have some errors, it is not clear how the fixed points of the kernel map will behave. In this section, we will consider stability as the property that if the pre-activations have some deviation from the zero mean, unit variance Gaussian, the post-activations will still converge to the fixed point of the kernel map. In other words, we would like to find out the following dynamics, have locally or globally atracting fixed points:

$$\mu_{\ell+1} = \mathbb{E}\ \phi(X),\ \sigma^2_{\ell+1} = \text{Var}(\phi(X)), \qquad\qquad X \sim N(\mu_\ell, \sigma^2_\ell), \tag{6.1}$$

where μ_0 and σ_0^2 are the mean and variance of the input. Note that if we have centered and properly scaled activation $\mathbb{E}\phi(X) = 0, \mathbb{E}\phi(X)^2 = 1$, then $\mu = 0, \sigma^2 = 1$ are the fixed points of the kernel map. However, in the finite width regime, we would like to know if the fixed points are locally or globally attractive. In other words, we would like to see if any purturbation from a fixed point will be contracted back towards the fixed point or deviate from it.

$$\mathrm{He}_n(x+y) = \sum_{k=0}^{n} \binom{n}{k} x^{n-k} \mathrm{He}_k(y) \tag{6.2}$$

$$\mathrm{He}_n(\gamma x) = \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \gamma^{n-2i}(\gamma^2-1)^i \binom{n}{2i} \frac{(2i)!}{i!} 2^{-i} \mathrm{He}_{n-2i}(x). \tag{6.3}$$

Using this property, we can prove the following local norm stability condition.

Lemma 6.1. *Let $X \sim N(0,1)$ and let ϕ be a function with Hermite expansion*

$$\phi(x) = \sum_{k=0}^{\infty} c_k \mathrm{He}_k(x),$$

Define $\gamma = 1+\epsilon$ for a small parameter ϵ. Then, up to first order in ϵ, the multiplicative purtorbation of norms $\mathbb{E}[\phi(\gamma X)^2]$ is given by

$$\mathbb{E}[\phi(\gamma X)^2] = \mathbb{E}[\phi(X)^2] + \epsilon \left(2\sum_{n=0}^{\infty} n!\, n\, c_n^2 + 4\sum_{n=0}^{\infty} n!\, c_n\, c_{n+2}\, (n+1)(n+2) \right) + \mathcal{O}(\epsilon^2).$$

and for the additive purturbation, we have

$$\mathbb{E}_{X\sim N(0,1)}\phi(X+\mu) = \sum_{n=0}^{\infty} c_n \mu^n$$

Note that the first order term in the multiplicative purturbation determines if the norm of the post-activations moves back towards to the fixed point, or moves away from it. In other words, the sign of the term $2\sum_{n=0}^{\infty} n! n c_n^2 + 4\sum_{n=0}^{\infty} n! c_n c_{n+2}(n+1)(n+2)$ will determine if the fixed point is locally attractive or repulsive. Based on this lemma, we can state the following sufficient condition for the local stability of the norm:

$$\frac{c_{n+2}}{c_n} < -\frac{2n}{(n+1)(n+2)}. \tag{6.4}$$

In other words, if the sign of coefficinets n and $n+2$ are opposite, and the magnitude of the ratio is approximately $2/n$ or larger, then the fixed point is locally attractive. Remarkably, this condition is met very accurately for SeLU, which was designed to be stable.

Proof. The proof of additive purturbation is straightforward, and follows directly from the formula for the addition. The rest of the proof focuses on the multiplicative purturbation. **Step 1: Hermite Expansion of $\phi(\gamma X)$**

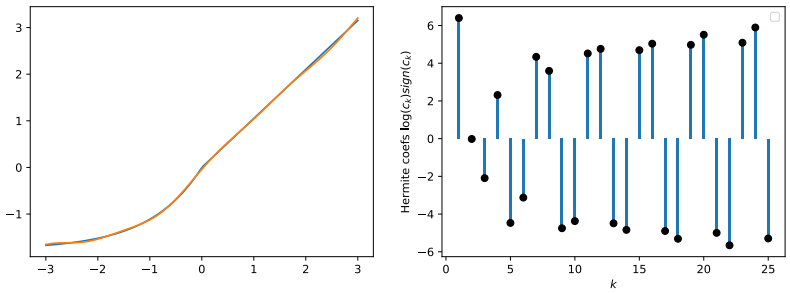

Figure 6.1: Stability of SELU activation, and its Hermite expansion

Given the Hermite expansion of ϕ,

$$\phi(x) = \sum_{k=0}^{\infty} c_k \, \mathrm{He}_k(x),$$

we can express $\phi(\gamma X)$ as

$$\phi(\gamma X) = \sum_{k=0}^{\infty} c_k \, \mathrm{He}_k(\gamma X).$$

Step 2: Apply the Multiplication Theorem

The multiplication theorem for Hermite polynomials states that for $\gamma = 1 + \epsilon$,

$$\mathrm{He}_n(\gamma x) = \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} \gamma^{n-2i} (\gamma^2 - 1)^i \binom{n}{2i} \frac{(2i)!}{i!} 2^{-i} \, \mathrm{He}_{n-2i}(x).$$

Substituting $\gamma = 1 + \epsilon$ and expanding up to first order in ϵ, we have

$$\gamma^n \approx 1 + n\epsilon, \quad \gamma^2 - 1 \approx 2\epsilon.$$

Thus,

$$\mathrm{He}_n(\gamma x) \approx \sum_{i=0}^{\lfloor \frac{n}{2} \rfloor} (1 + n\epsilon)^{n-2i} \, (2\epsilon)^i \binom{n}{2i} \frac{(2i)!}{i!} 2^{-i} \, \mathrm{He}_{n-2i}(x).$$

To first order in ϵ,

$$\mathrm{He}_n(\gamma x) \approx \mathrm{He}_n(x) + 2i\epsilon \cdot (\text{coefficients}).$$

Step 3: Expand $\phi(\gamma X)$ and Compute $\mathbb{E}[\phi(\gamma X)^2]$

Expanding $\phi(\gamma X)$ using the above approximation:

$$\phi(\gamma X) \approx \sum_{k=0}^{\infty} c_k \left[\operatorname{He}_k(x) + \epsilon \cdot (\text{linear terms in } \operatorname{He}_{k-2i}(x)) \right].$$

When squaring $\phi(\gamma X)$ and taking expectation, due to the orthogonality of Hermite polynomials:

$$\mathbb{E}[\operatorname{He}_n(X) \operatorname{He}_m(X)] = n! \delta_{n,m},$$

where $\delta_{n,m}$ is the Kronecker delta.

Therefore,

$$\mathbb{E}[\phi(\gamma X)^2] = \sum_{n=0}^{\infty} n!\, d_n^2,$$

where d_n are the coefficients in the Hermite expansion of $\phi(\gamma X)$. Expanding d_n up to first order in ϵ, we obtain

$$d_n \approx c_n(1 + n\epsilon) + 2\epsilon\, c_{n+2}(n+1)(n+2).$$

Squaring and keeping terms up to first order in ϵ:

$$d_n^2 \approx c_n^2(1 + 2n\epsilon) + 4\epsilon\, c_n c_{n+2}(n+1)(n+2).$$

Thus,

$$\mathbb{E}[\phi(\gamma X)^2] \approx \sum_{n=0}^{\infty} n! \left(c_n^2 + 2n c_n^2 \epsilon + 4 c_n c_{n+2}(n+1)(n+2)\epsilon \right).$$

This simplifies to

$$\mathbb{E}[\phi(\gamma X)^2] = \mathbb{E}[\phi(X)^2] + \epsilon \left(2 \sum_{n=0}^{\infty} n!\, n\, c_n^2 + 4 \sum_{n=0}^{\infty} n!\, c_n\, c_{n+2}\, (n+1)(n+2) \right) + \mathcal{O}(\epsilon^2),$$

where $\mathbb{E}[\phi(X)^2] = \sum_{n=0}^{\infty} n!\, c_n^2$.

□

6.2 Gradients and neural tangent kernel

So far, we only considered the forward passes. In this section, we consider the backward gradients, which will demosntate the importance of the convergence results that was established in the previous sections. Before that, let us study a few elegant properties of Hermite polynomials that will be useful later.

Lemma 6.2. *Let ϕ be an activation with Hermite expansion $\phi(z) = \sum_{k=0}^{\infty} c_k \operatorname{he}(z)$, and with kernel map κ. Then, its derivative has Hermite expansion*

$$\phi'(z) = \sum_{k=1}^{\infty} \sqrt{k} c_k \operatorname{he}_{k-1}(z). \tag{6.5}$$

Furthermore, if X, Y are standard Gaussians with covariance ρ, the kernel map of the derivative is given by

$$\mathbb{E}\phi'(X)\phi'(Y) = \kappa'(\rho) = \sum_{k=1}^{\infty} k c_k^2 \rho^{k-1}. \tag{6.6}$$

This lemma shows a remarkable property of Hermite polynomilas, in that, the kernel map of derivative, is derivative of the kernel map of the original function. The main proof step is the fact that Hermite polynomials constitute an Appell sequence, i.e., $\operatorname{He}_k'(z) = \sqrt{k} \operatorname{He}_{k-1}(z)$.

Proof. The derivative of probablist's Hermite polynomials is directly linked tho the lower degree Hermite polynomials, allowing us to write

$$\begin{aligned} \frac{d}{dz} \operatorname{He}_k'(z) &= k \operatorname{He}_{k-1}(z) \\ \implies \operatorname{he}_k'(z) &= \sqrt{k} \operatorname{he}_{k-1}(z) \\ \implies \phi'(z) &= \sum_{k=1}^{\infty} \sqrt{k} c_k \operatorname{he}_{k-1}(z) \end{aligned}$$

Thus, we can conclude the proof by invoking Lemma 5.6

$$\begin{aligned} \mathbb{E}\phi'(X)\phi'(Y) &= \mathbb{E}\left[\sum_{k=1}^{\infty} \sqrt{k} c_k \operatorname{he}_{k-1}(X) \sum_{k=1}^{\infty} \sqrt{k} c_k \operatorname{he}_{k-1}(Y)\right] \\ &= \sum_{k=1,n=1}^{\infty} \sqrt{k} c_k \sqrt{n} c_n \mathbb{E}[\operatorname{he}_{k-1}(X) \operatorname{he}_{n-1}(Y)] \\ &= \sum_{k=1}^{\infty} k c_k^2 \mathbb{E}[\operatorname{he}_{k-1}(X) \operatorname{he}_{k-1}(Y)]^{k-1} && \text{Lemma 5.6} \\ &= \sum_{k=1}^{\infty} k c_k^2 \rho^{k-1} = \kappa'(\rho) \end{aligned}$$

□

Let us now consider the gradients of the MLPs

6.2.1 Gradients of MLPs

Given input dimension d_0, output dimension d_L, and hidden dimensions $d_1, \dots, d_{L-1}$, input $x \ni \mathbb{R}^{d_0}$ and output $y \in \mathbb{R}^{d_L}$, the MLP is defined as

$$\begin{aligned}
&h^0 := x, && x \in \mathbb{R}^{d_0} \\
&z^\ell := W_\ell h^{\ell-1}, && W_\ell \in \mathbb{R}^{d_\ell \times d_{\ell-1}}, \text{drawn iid from } \sim N(0, 1/d) \\
&h^\ell := \phi(z^\ell), && \ell = 1, \dots, L \\
&\hat{y} := z^L = W_L h^{L-1}, && \hat{y}, y \in \mathbb{R}^{d_L}
\end{aligned}$$

Assuming sum squared, we can deinfe

$$\mathcal{L}(x, y) = \frac{1}{2}\|\hat{y} - y\|^2,$$

the gradients of the last layer are given by

$$\begin{aligned}
\delta^L &:= \frac{\partial \mathcal{L}}{\partial \hat{y}} = \hat{y} - y \\
\delta^\ell &:= \frac{\partial \mathcal{L}}{\partial z^\ell} = \phi'(z^\ell) \odot (W_{\ell+1}^\top \delta^{\ell+1}), && \ell = 1, \dots, L-1 \\
g^\ell &:= \frac{\partial \mathcal{L}}{\partial W_\ell} = \delta^\ell \otimes h^{\ell-1}, && \ell = 1, \dots, L,
\end{aligned}$$

where $\otimes$ denotes the outer product, and the dependece of $\mathcal{L}$ and $\hat{y}$ on x is omitted for brevity.

The main goal of this section is to analyze the norm of the gradients in the mean field regime, and then to exend this to the Neural Tangent Kernel (NTK). In other words, we are interested in properties of $\|g^l\|_{rms}$, when $d_1, \dots, d_{L-1} \to \infty$, and wether it vanishes or explodes with the depth of the network. For the NTK, we are interested in the properties of the sequence $\langle g^l, g^l \rangle$, when $d_1, \dots, d_{L-1} \to \infty$.

In the remainder of this section, we will use $\rho_{X,Y}$ to refer to average inner product $\rho_{X,Y} = \frac{1}{n}\sum_{i=1}^n X_i^\top Y_i$. Furthermore, we will use $\|X\|_{rms}$ to refer to the RMS norm $\frac{1}{n}\sum_{i=1}^n x_i^2$.

Proposition 6.3. *Let ϕ be an activation that obeys $\mathbb{E}_{X \sim N(0,1)} \phi(X)^2 = 1$. In the mean field regime, the norm of the gradients obeys*

$$\|\delta^\ell\|_{rms} = \|\delta^{\ell+1}\|_{rms} \kappa'(1) \tag{6.7}$$

Furthermore, for two inputs with initial similarity $\rho_0 = \langle x_1, x_2\rangle$ the forward and backward passes follow recursion

$$\rho_\ell := \rho_{h_1^\ell, h_2^\ell} = \kappa(\rho_{\ell-1}) \tag{6.8}$$

$$\langle \delta_1^\ell, \delta_2^\ell\rangle = \kappa'(\rho^\ell)\langle \delta_1^{\ell+1}, \delta_2^{\ell+1}\rangle \tag{6.9}$$

$$\langle g_1^\ell, g_2^\ell\rangle = \rho^{\ell-1}\langle \delta_1^\ell, \delta_2^\ell\rangle. \tag{6.10}$$

Let us consider the backward errors δ^l and their norms. Each index of δ^l is a Gaussian weights at layer $l+1$, by δ^{l+1} and $\phi'(z^l)$. Conditioned on the norm of δ^{l+1}, with the independece of weights of at layer $l+1$ and pre-activations at layer l, norm of δ^l is the multiplication of norm of δ^{l+1} and the norm of $\phi'(z^l)$. Now, assuming that the activation is centered, pre-activations are standard Gaussian, and thus $\mathbb{E}\phi'(z^l)^2$ can be expanded based on the Hermite expansion and properties of Hermite polynomials

$$\begin{aligned}
\phi'(z) &= \sum_{k=1}^{\infty} c_k \,\mathrm{he}_{k-1}(z) \\
\mathbb{E}\phi'(z^l)^2 &= \sum_{k=1}^{\infty} k c c_k^2 =: \beta \\
\|\delta^l\|_{rms} &= \|\delta^{l+1}\|_{rms}\mathbb{E}\phi'(z^l)^2 = \|\delta^{l+1}\|_{rms}\beta \\
\|\delta^l\|_{rms} &= \|\delta^L\|_{rms}\beta^{L-l} = \|y-\hat{y}\|_{rms}\beta^{L-l}
\end{aligned}$$

Remark 6.4. Note that assuming that the activation unit second momnt $\mathbb{E}\phi(X)^2 = 1$, we have $\sum_{k=1}^{\infty} c_k^2 = 1$, which implies that $\beta \leq 1$. Furthermore, if the activation is non-linear, i.e., there is $c_k \neq 0$ for some $k \geq 2$, then the inequality is strict, i.e., $\beta < 1$. Recall that for the forward pass of an activation that centered and unit variance, we have shown the the kernel sequence converges towards zero with rate $\alpha = 1/(\sum_{k=1} c_k^2)$, where the rate becomes strict whenever we have any non-linearity. This implies that the stronger the non-linearity, the faster the convergence of kernel sequence towards zero, and the faster the gradients norms converge towards zero. This shows that striking a balance between linearity and non-linearity is probably needed for a desirable initialization.

Now, consider a generic layer $\delta' = \mathbf{W}^\top \delta \odot \phi'(\mathbf{z})$. Suppose z_1, z_2 correspond to pre-activations for two different inputs, and δ_1, δ_2 are the corresponding gradients for previous and δ_1', δ_2' are the gradients for the next layer.

If we look at one corresponding unit of $W_\top \delta_1$ and $W_\top \delta_2$, they are jointly Gaussian with covariance $\mathbb{E}\delta_1\delta_2$. Thus, in the mean field, we can write

$$\mathbb{E}\delta_1'\delta_2' = (\mathbb{E}\delta_1\delta_2)(\mathbb{E}\phi'(z_1)\phi'(z_2)) = \mathbb{E}\delta_1\delta_2 f(\rho),$$

where we used the fact that error-propagation vectors $\delta_1\delta_2$ are independent of pre-activations $z_1 z_2$, because one relies on layers before and the other on layers after this step.

Now, in a multi-layer setting, let $r_l = \frac{1}{d}\langle \delta_1^l, \delta_2^l \rangle$ denote the normalized inner product of the gradients at layer l, and let $\rho_\ell = \frac{1}{d}\langle z_1^l, z_2^l \rangle$ denote the normalized inner product of the pre-activations at layer ℓ. Then, in the mean field regime, we have

$$r_l = f(\rho_l) r_{l+1}$$

Thus, for the weight gradients of layer ℓ, we have