Notes on Neural Tangent Kernel: Convergence and Generalization in Neural Networks, by Jacot et al.

for the Fall2019 NTK reading group in Oxford

Guillermo Valle Pérez

August 19, 2019

These are notes explaining the results and proofs in the paper introducing Neural Tangent Kernels, by Jacot et al. [2]. It studies a limiting case of gradient descent-trained neural networks, which is analytically tractable, and may offer theoretical insight into why neural networks work, in some cases.

It is clear that, when training a neural network, if the learning rate is made very small, and the training time is kept fixed, then we are effectively fitting a linear model, given by the first term in the Taylor expansion of the network's function with respect to the network's parameters, centered around the parameters at initialization. However, for a fixed network architecture, the total change in the network's function after training also becomes smaller and smaller amount, as we make the learning rate smaller. This clearly limits the use of such neural network, as it can't change the functions significantly, and thus fit the training data.

Jacot's paper shows that if we make the network wider at the same time as we decrease the learning rate, then the network still approaches its linearization (uniformly within a fixed training time), but the network's function can change by an $O(1)$ amount. In fact in the limit, the network's function evolves according to a "kernel gradient descent" equation with a well-defined limiting kernel, known as the *neural tangent kernel* (NTK). This means that this is a "useful limit" because the network can fit data (which realistically requires $O(1)$ in the function), but is also analytically tractable because it is a linear model. Furthermore, they show that in the infinite width limit, the NTK becomes independent of initialization, which further simplifies the study of the network.

To prove that the linear model can fit *any* data requires the NTK to be positive definite. This is an important step in several papers studying convergence of wide networks with small learning rates. In Jacot et al., they also prove positive definitness under some simplifying assumptions.

The paper, however, doesn't discuss whether this limit describes the behavior of realistic neural networks, which are not infinitely wide, and don't have infinitesimal learning rates. This question has been studied in subsequent literature. Their results also require smooth activation functions, which don't include the most popular one, the ReLU. Few works have attempted to obtain similar results to include this case, but it makes the analysis much more complicated.

1 Notation

We consider fully-connected feedforward neural networks. We call "the number of layers" the number affine transformations, so that the number of "hidden layers" is one less than the number of layers. We will typically refer to hidden layers as "layers". Hopefully the context, and the pictures makes it clear.

For layer l, preactivations are written $\tilde{\alpha}^{(l)}$, activations $\alpha^{(l)}$, weights $W^{(l)}$, and biases $b^{(l)}$. See fig. 1. The width of (number of neurons in) the lth layer is n_l. The 0th "activations" are just the inputs of the networks, $\alpha^{(0)} = x$.

We often call the output of the last layer $f(x) = \tilde{\alpha}^{(L)}(x)$ (for a network with L layers). The gradient of the cost function with respect to the network outputs is called d.

We define a data-dependent inner product for functions on the input space, that is functions in $\mathcal{F} = \{f : \mathbb{R}^{n_0} \to \mathbb{R}^k\}$ for some k. For $f, g \in \mathcal{F}$, we define $\langle f, g \rangle_{p^{\text{in}}} := \mathbb{E}_{x \sim p^{\text{in}}} \left[f^T(x) g(x) \right]$, where p^{in} is a distribution over inputs. Throughout these notes, we implicitly assume that p^{in} is the empirical distribution (uniformly sampling inputs in the training dataset), although results can typically be extended to general distributions.

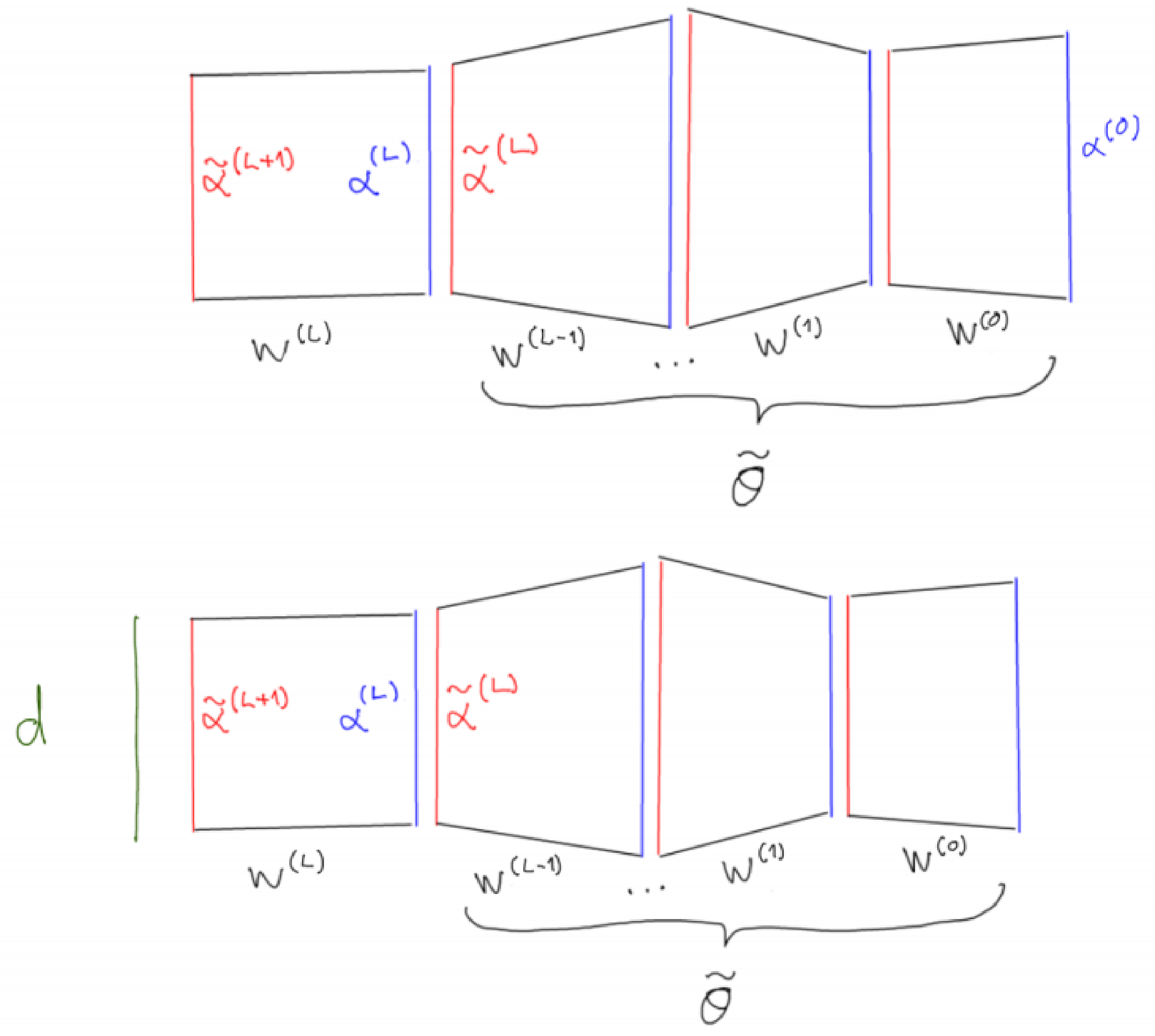


Figure 1: A network with $L+1$ layers (remember that we count layers as the number of affine transformations, so the number of hidden layers here is L, but the number of *layers* is $L+1$. We show here a network with $L+1$ layers, so that this figure can be used for quick reference when we do inductive proofs (which we'll do a lot). $\tilde{\theta}$ is the name given to all the parameters *except* those in the last layer. This is used in some proofs.

I will sometimes use Einstein's summation convention, whereby if two indices are repeated in a term, they are implicitly summed over.

2 Neural tangent kernel

The definition of the neural tangent kernel (NTK) $\Theta_{kk'}(x, x') = \Theta^{(L)}_{kk'}(x, x')$ for a network with parameters θ and L layers, is

$$\Theta^{(L)}_{kk'}(x, x') = \sum_{\theta} \partial_\theta f_k(x)\, \partial_\theta f_{k'}(x')$$

We also define $\Theta^{(l)}(x, x')$, the NTK at the lth ($l < L$) layer of a neural network of L layers, by substituting the outputs f in the definition above, by $\tilde{\alpha}^{(l)}$, the preactivations at layer l.

This matrix comes out from the equation describing the dynamics of f, the vector of outputs, which I derive below. Here I am interpreting f as the vector of all outputs over all inputs in a finite training set. For each input x_i, $i = 1, ..., N$, there are n_L outputs of the network ($f_k(x_i)$ $k = 1, ..., n_L$). Concatenate all of these ouputs to obtain the Nn_L-dimensional vector f. This is less general than what Jacot considers (they consider f to be a function over a potentially infinite input domain), but I think it captures the idea with simpler notation/language. My approach here is the same as that in Lee et al. [3] (in https://arxiv.org/abs/1902.06720).

2.1 Kernel gradient

Here is the derivation of the kernel gradient flow $\partial_t f$, derived from the gradient flow of parameters $\partial_t \theta$.

KERNEL GRADIENT

GD: $\partial_t \theta_j = - \partial_{f_k} C \; \partial_{\theta_j} f_k$

$$\partial_t f_i = \partial_{\theta_j} f_i \; \partial_t \theta_j = \underbrace{\partial_{\theta_j} f_i \; \partial_{\theta_j} f_k}_{\Theta_{ik}} \; \underbrace{\partial_{f_k} C}_{d_k}$$

↑ FIXED FOR LINEAR MODELS

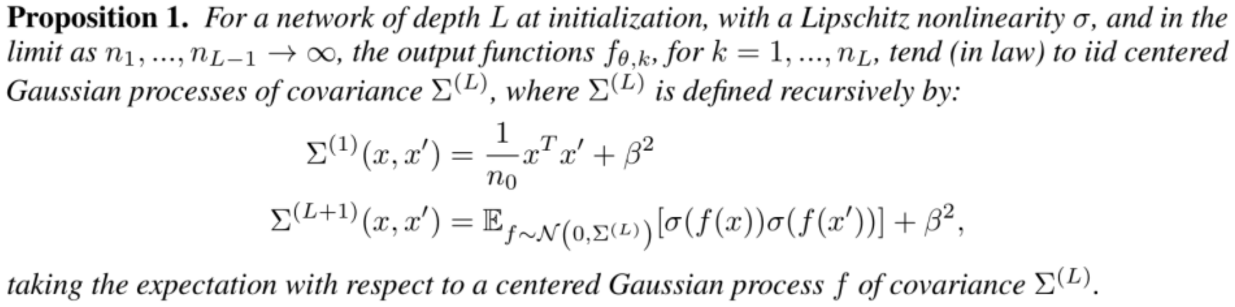
2.2 NTK parametrization

Furthermore in this paper they introduce the *NTK parametrization*, whereby each weight W^L gets substituted by $\frac{1}{n_L} W^{(L)}$. While in the original parametrization, the weights had variance $O(1/\sqrt{n_L})$, now they are taken to have variance $O(1)$. The network function stays the same, but the parameters that parametrize it are different.

3 Gaussian process ∞ width limit

In the limit of infinite width $n_{1,\ldots,L} \to \infty$, the vector of outputs f is distributed according to a Gaussian distribution with a block diagonal covariance matrix where two outputs for the same input x, $f_k(x)$ and $f_{k'}(x)$ have 0 correlation. As this is true for any finite set of inputs, this is (almost) enough to show that the full function f_θ is sampled from a Gaussian process. However, they don't deal with the technical issues of showing that, and instead focus on finite number of inputs for simplicity, in this case.

The precise proposition (which gives a recursive formula for the covariance matrix) is:



Proposition 1. *For a network of depth L at initialization, with a Lipschitz nonlinearity σ, and in the limit as $n_1, ..., n_{L-1} \to \infty$, the output functions $f_{\theta,k}$, for $k = 1, ..., n_L$, tend (in law) to iid centered Gaussian processes of covariance $\Sigma^{(L)}$, where $\Sigma^{(L)}$ is defined recursively by:*

$$\Sigma^{(1)}(x, x') = \frac{1}{n_0} x^T x' + \beta^2$$
$$\Sigma^{(L+1)}(x, x') = \mathbb{E}_{f \sim \mathcal{N}\left(0, \Sigma^{(L)}\right)}[\sigma(f(x))\sigma(f(x'))] + \beta^2,$$

taking the expectation with respect to a centered Gaussian process f of covariance $\Sigma^{(L)}$.

Note that instead of talking about a block diagonal covariance of the full f, they just derive the covariance between $f_k(x)$ and $f_k(x')$ for the N different inputs, for a one output k, and say that different outputs k are i.i.d., which is equivalent to saying that the full Nn_L-dimensional vector f is block diagonal with blocks of size N.

Prove this by induction.

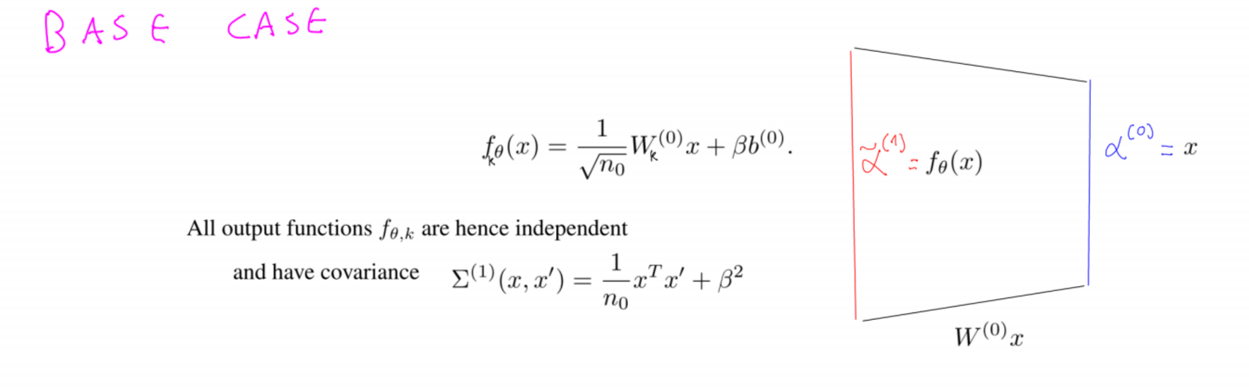
3.1 Base case

The base case simply writes $\alpha^{(1)}$ (which can be thought of as a Nn_1-dim vector) as linear combination of linear transformations of a Gaussian vector, which is still Gaussian:

$$\tilde{\alpha}_k^{(1)}(x_i) = X_{il} w_{kl} + \beta b_k$$

w and b being $W^{(0)}$ and $b^{(0)}$ respectively, X is the data matrix with rows as data points, and $i = 1, .., N$, $k = 1, .., n_1$. Therefore, as they are jointly Gaussian, one just then needs to show that $\alpha_k^{(1)}(x)$ and $\alpha_{k'}^{(1)}(x')$ have zero correlation to show that they are independent. This follows immediately from the fact that different weights are independent so that $\mathbf{E}[w_{kl} w_{k'l'}] = \delta_{kk'} \delta_{ll'}$. In more detail the covariance between two elements of $\alpha^{(1)}$ is

$$\begin{aligned}
\mathbf{E}[\tilde{\alpha}_k^{(1)}(x_i) \tilde{\alpha}_{k'}^{(1)}(x_{i'})] &= \frac{1}{n_0} X_{il} \mathbf{E}[w_{kl} w_{k'l'}] X_{i'l'} + \beta^2 \mathbf{E}[b_k b_{k'}] \\
&= \frac{1}{n_0} X_{il} X_{i'l} \delta_{kk'} + \beta^2 \delta_{kk'} \\
&\equiv \delta_{kk'} \Sigma^{(1)}(x_i, x_{i'})
\end{aligned}$$



BASE CASE

$$f_{\theta_k}(x) = \frac{1}{\sqrt{n_0}} W_k^{(0)} x + \beta b^{(0)}.$$

All output functions $f_{\theta,k}$ are hence independent

and have covariance $\Sigma^{(1)}(x,x') = \frac{1}{n_0} x^T x' + \beta^2$

3.2 inductive step

The inductive hypothesis is that the preactivations of a layer $\tilde{\alpha}_k^{(L)}(x)$ are i.i.d. with respect to the index k (that is the preactivations of different neurons in the Lth layer are i.i.d.). The activations are just elementwise functios of the preactivations and so are also i.i.d. w.r.t. k. Write the preacts of the next layer as

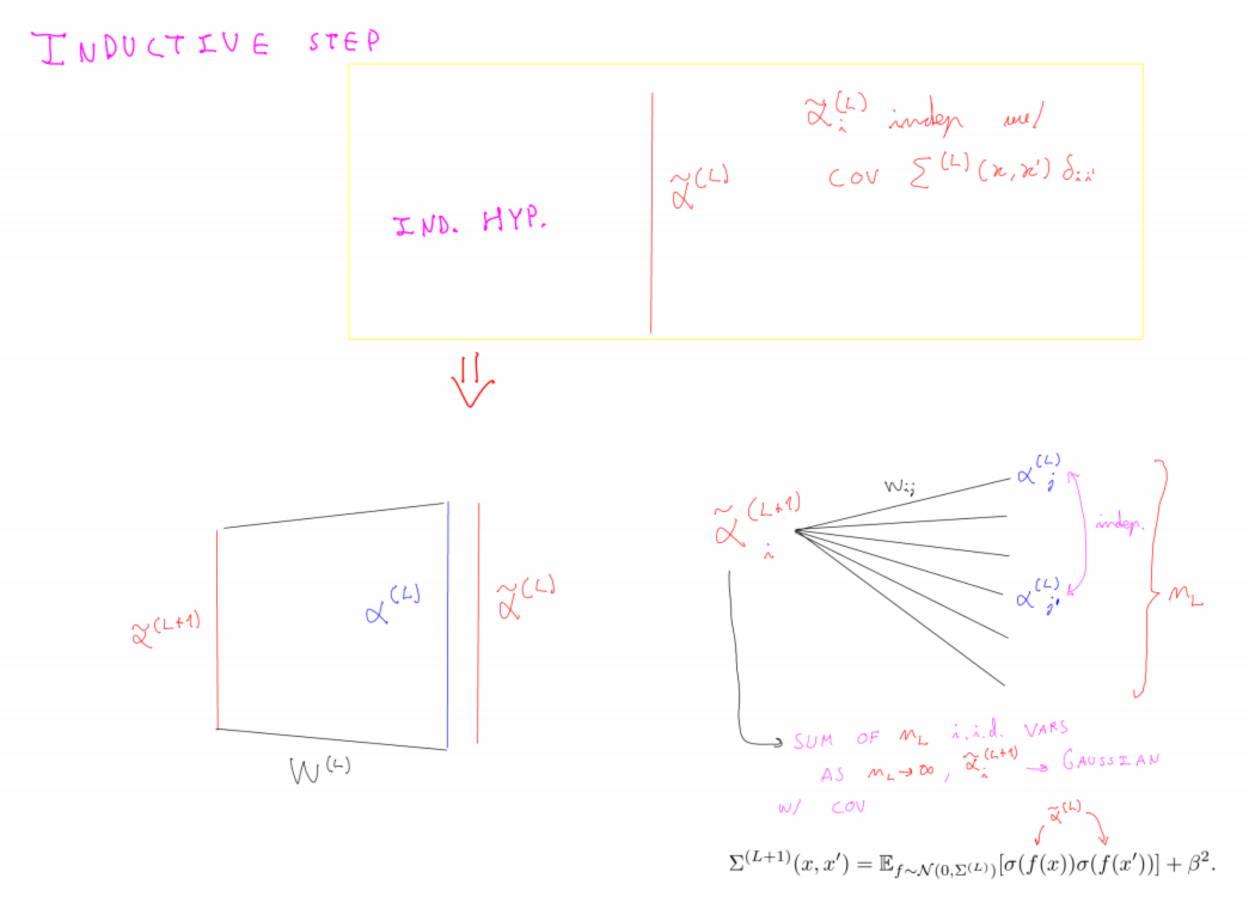
$$\tilde{\alpha}_k^{(L+1)}(x_i) = \sum_{l=1}^{n_L} \alpha_l^{(L)}(x_i) w_{kl} + \beta b_k \tag{1}$$

w and b now being $W^{(L)}$ and $b^{(L)}$ respectively, X is the data matrix with rows as data points, and $i = 1,..,N$, $k = 1,..,n_1$. Therefore, the Nn_{L+1}-dim vector $\alpha^{(L+1)}$ is a sum of n_L i.i.d. vectors. Let me try to make that more clear, by writting it explicitly as Nn_{L+1}-dim vectors

$$\begin{pmatrix} \tilde{\alpha}_l^{(L+1)}(x_1) \\ \tilde{\alpha}_l^{(L+1)}(x_2) \\ \vdots \\ \tilde{\alpha}_l^{(L+1)}(x_N) \end{pmatrix} = \sum_{l=1}^{n_L} \begin{pmatrix} \alpha_l^{(L)}(x_1) w_{1l} \\ \alpha_l^{(L)}(x_2) w_{1l} \\ \vdots \\ \alpha_l^{(L)}(x_N) w_{n_{L+1}l} \end{pmatrix} + \begin{pmatrix} b_1 \\ b_1 \\ \vdots \\ b_{n_{L+1}} \end{pmatrix}$$

Because each of the vectors in the sum is i.i.d. (because all the weights are, and the $\tilde{\alpha}_k^{(L)}$ were assumed to be w.r.t. k by the inductive hypothesis), then when $n_L \to \infty$, one can use the multidimensional central limit theorem, to say that the sum becomes a multidimensional Gaussian. The vector of biases is Gaussian at any width anyways. Therefore their sum is Gaussian as $n_L \to \infty$.

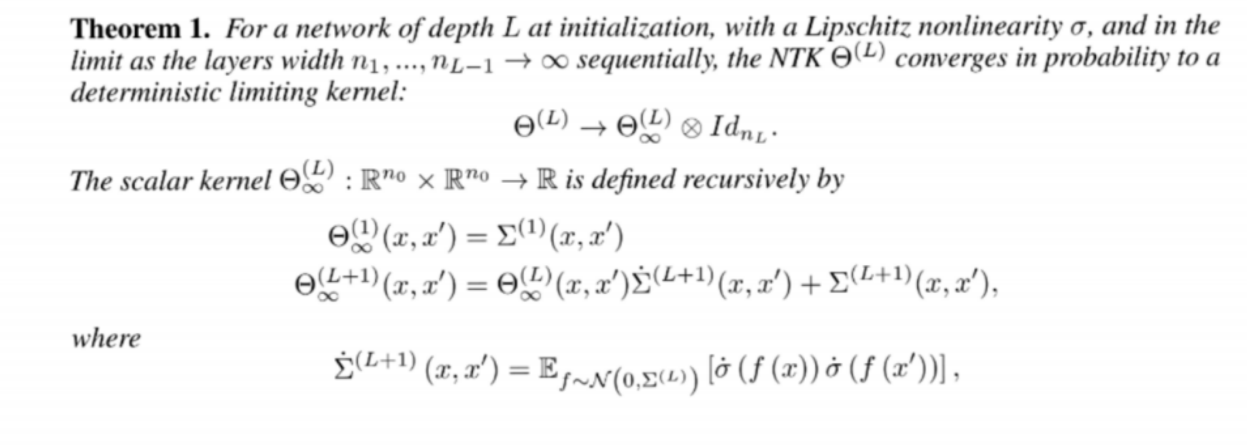
Again, all that's left to show is that the $\tilde{\alpha}_k^{(L+1)}(x)$ are i.i.d. w.r.t. k. As we have now shown they are jointly Gaussian, this can be shown by computing the covariance, and again finding that $\mathbf{E}[\tilde{\alpha}_k^{(L+1)}(x)\tilde{\alpha}_k^{(L+1)}(x')]$ is independent of k (identically distributed) and that $\mathbf{E}[\tilde{\alpha}_k^{(L+1)}(x)\tilde{\alpha}_{k'}^{(L+1)}(x')] = 0$ when $k \neq k'$ (independent). This is easy to compute using the Equation 1, very similarly to how we did in the base case. The answer is $\delta_{kk'}\Sigma^{(L+1)}(x,x')$ (defined on statement of propostion, and in illustration below).



$$\Sigma^{(L+1)}(x,x') = \mathbb{E}_{f\sim\mathcal{N}(0,\Sigma^{(L)})}[\sigma(f(x))\sigma(f(x'))] + \beta^2.$$

4 NTK determinitic ∞-width limit

We show that the NTK stochastically converges to a deterministic limit as the width becomes infinite. This means that in the limit, almost surely (with probability approaching 1), at initialization, the NTK gets arbitrarily close to a certain limit. For finite width networks, the NTK would depend on the particular parameters we get at initialization.



Theorem 1. *For a network of depth L at initialization, with a Lipschitz nonlinearity σ, and in the limit as the layers width $n_1, ..., n_{L-1} \to \infty$ sequentially, the NTK $\Theta^{(L)}$ converges in probability to a deterministic limiting kernel:*

$$\Theta^{(L)} \to \Theta_\infty^{(L)} \otimes Id_{n_L}.$$

The scalar kernel $\Theta_\infty^{(L)} : \mathbb{R}^{n_0} \times \mathbb{R}^{n_0} \to \mathbb{R}$ is defined recursively by

$$\Theta_\infty^{(1)}(x, x') = \Sigma^{(1)}(x, x')$$
$$\Theta_\infty^{(L+1)}(x, x') = \Theta_\infty^{(L)}(x, x')\dot{\Sigma}^{(L+1)}(x, x') + \Sigma^{(L+1)}(x, x'),$$

where

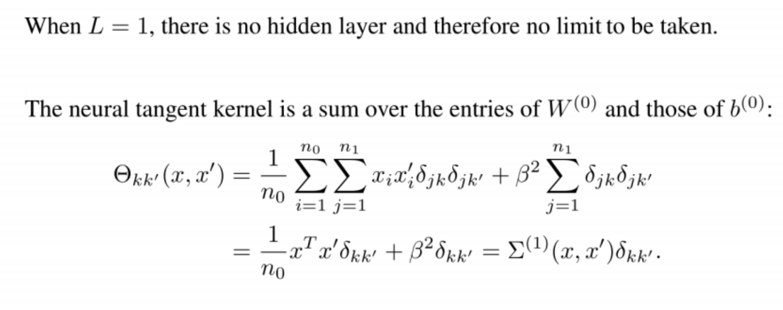
$$\dot{\Sigma}^{(L+1)}(x, x') = \mathbb{E}_{f \sim \mathcal{N}(0, \Sigma^{(L)})}\left[\dot{\sigma}(f(x))\,\dot{\sigma}(f(x'))\right],$$

Note that $\Sigma^{(L)}(x, x')$ here is the kernel describing the distribution of activations at initialization in the infinite width limit, as derived in the previous section.

We again prove this by induction

4.1 Base case

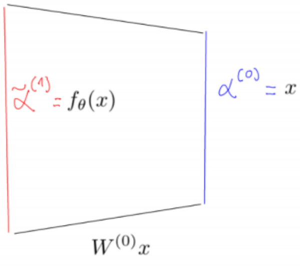
The base case just comes from differentiating $\tilde{\alpha}^{(1)}$ with respect to $W^{(0)}$ and $b^{(0)}$



When $L = 1$, there is no hidden layer and therefore no limit to be taken.

The neural tangent kernel is a sum over the entries of $W^{(0)}$ and those of $b^{(0)}$:

$$\begin{aligned}\Theta_{kk'}(x, x') &= \frac{1}{n_0}\sum_{i=1}^{n_0}\sum_{j=1}^{n_1} x_i x_i' \delta_{jk}\delta_{jk'} + \beta^2 \sum_{j=1}^{n_1} \delta_{jk}\delta_{jk'} \\ &= \frac{1}{n_0} x^T x' \delta_{kk'} + \beta^2 \delta_{kk'} = \Sigma^{(1)}(x, x')\delta_{kk'}.\end{aligned}$$

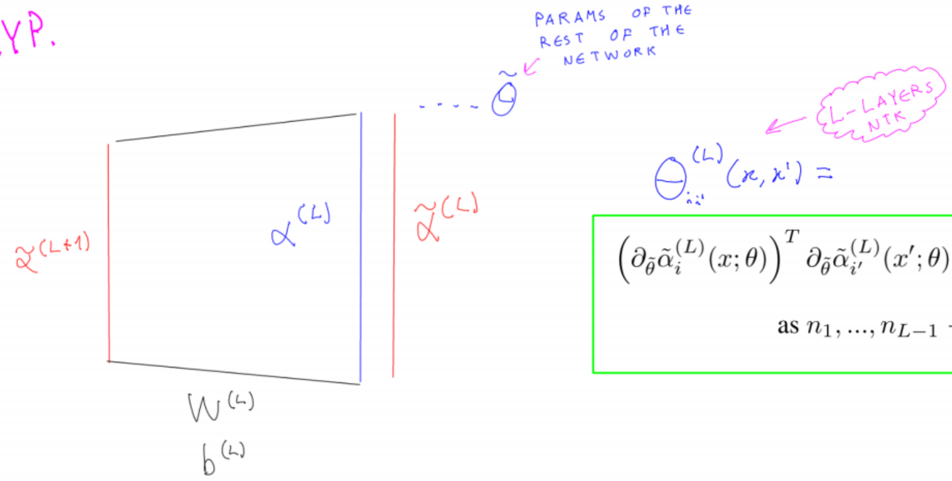


This gives us that $\Theta_\infty^{(1)}(x, x') = \Sigma^{(1)}(x, x')$

4.2 Inductive step

The inductive hypothesis is that, the theorem is true for L layers. We consider the behavior of the activations at the $L + 1$ th layer.

IND. HYP.



$$\Theta^{(L)}_{ii'}(x,x') =$$

$$\boxed{\begin{gathered}\left(\partial_{\tilde{\theta}}\tilde{\alpha}_i^{(L)}(x;\theta)\right)^T \partial_{\tilde{\theta}}\tilde{\alpha}_{i'}^{(L)}(x';\theta) \to \Theta_\infty^{(L)}(x,x')\delta_{ii'}.\\ \text{as } n_1, ..., n_{L-1} \to \infty:\end{gathered}}$$

To prove the inductive step, we split the NTK at the $L+1$ th layer into a sum of derivatives w.r.t. the last $(L+1)$th layer parameters, and w.r.t. the rest of the parameters which we collectively call $\tilde{\theta}$

SPLIT NTK

$$\Theta^{(L+1)}_{k,k'}(x,x') = \sum_{\substack{\theta \\ \text{LAST} \\ \text{LAYER}}} \partial_\theta f(x)\, \partial_\theta f(x') + \sum_{\substack{\tilde{\theta} \\ \text{REST OF} \\ \text{NETWORK}}} \partial_{\tilde{\theta}} f(x)\, \partial_{\tilde{\theta}} f(x')$$

CONTRIBUTION FROM LAST LAYER

$$= \sum_i \underbrace{\partial_{b_i^{(L)}} f_k(x)}_{\delta_{ik}} \underbrace{\partial_{b_i} f_{k'}(x')}_{\delta_{ik'}} = \beta^2 \delta_{kk'}$$

BIASES

$$+ \sum_{ij} \underbrace{\partial_{W_{ij}^{(L)}} f_k(x)}_{\frac{1}{\sqrt{m_L}}\delta_{jk}\alpha_i(x)} \underbrace{\partial_{W_{ij}^{(L)}} f(x')}_{\cdots} = \frac{1}{m_L}\delta_{kk'} \sum_i \alpha_i(x)\,\alpha_i(x')$$

WEIGHTS

$$\left.\right\} \Sigma^{(L+1)}(x,x')$$

$$+ \sum_{\tilde{\theta}_p} \partial_{\tilde{\theta}_p} f_{\theta,k}(x)\, \partial_{\tilde{\theta}_p} f_{\theta,k'}(x')$$

As we see the contribution from the last layer has a form analogous to the base case, and just gives us $\Sigma^{(L+1)}(x,x')$.

To compute the contribution from the rest of parameters, we first apply the chain rule to compute $\partial_{\tilde{\theta}_p} f_{\theta,k}(x)$ in terms of $\partial_{\tilde{\theta}_p} \tilde{\alpha}_i^{(L)}(x;\theta)$ which we understand from the inductive hypothesis, which we therefore apply. Finally, we use the resul from Proposition 1 (the initialization GP kernel limit), and the law of large numbers[1].

FOR $$\sum_{\underset{\text{REST OF NETWORK}}{\tilde{\Theta}}} \partial_{\tilde{\Theta}} f(x)\, \partial_{\tilde{\Theta}} f(x')$$

by the chain rule: $$\partial_{\tilde{\theta}_p} f_{\theta,k}(x) = \frac{1}{\sqrt{n_L}} \sum_{i=1}^{n_L} \partial_{\tilde{\theta}_p} \tilde{\alpha}_i^{(L)}(x;\theta) \dot{\sigma}(\tilde{\alpha}_i^{(L)}(x;\theta)) W_{ik}^{(L)}.$$

IND. HYP. $\Big($as $n_1, \ldots, n_{L-1} \to \infty$: $\Theta_{ii'}^{(L)}(x,x') \to \Theta_{\infty}^{(L)}(x,x')\delta_{ii'}.\Big)$

$$\Rightarrow \quad \frac{1}{n_L} \sum_{i,i'=1}^{n_L} \Theta_{ii'}^{(L)}(x,x') \dot{\sigma}\left(\tilde{\alpha}_i^{(L)}(x;\theta)\right) \dot{\sigma}\left(\tilde{\alpha}_{i'}^{(L)}(x';\theta)\right) W_{ik}^{(L)} W_{i'k'}^{(L)}$$

$$\to \frac{1}{n_L} \sum_{i=1}^{n_L} \Theta_{\infty}^{(L)}(x,x') \dot{\sigma}\left(\tilde{\alpha}_i^{(L)}(x;\theta)\right) \dot{\sigma}\left(\tilde{\alpha}_i^{(L)}(x';\theta)\right) W_{ik}^{(L)} W_{ik'}^{(L)}$$

BY PROP. 1, $\tilde{\alpha}_i^{(L)}(x;\theta)$ ARE i.i.d.

$\Downarrow$

By the law of large numbers, as $n_L \to \infty$, this tends to its expectation which is equal to

$$\Theta_{\infty}^{(L)}(x,x')\, \mathbb{E}_{f \sim \mathcal{N}\left(0,\Sigma^{(L)}\right)} \left[\dot{\sigma}\left(f\left(x\right)\right) \dot{\sigma}\left(f\left(x'\right)\right)\right]$$

$$= \Theta_{\infty}^{(L)}(x,x') \dot{\Sigma}^{(L+1)}(x,x')\delta_{kk'}.$$

□

[1] We could alternatively apply the Central Limit theorem if we wanted, to get that the sum approaches a Guassian with vanishing variance

5 NTK constant during training

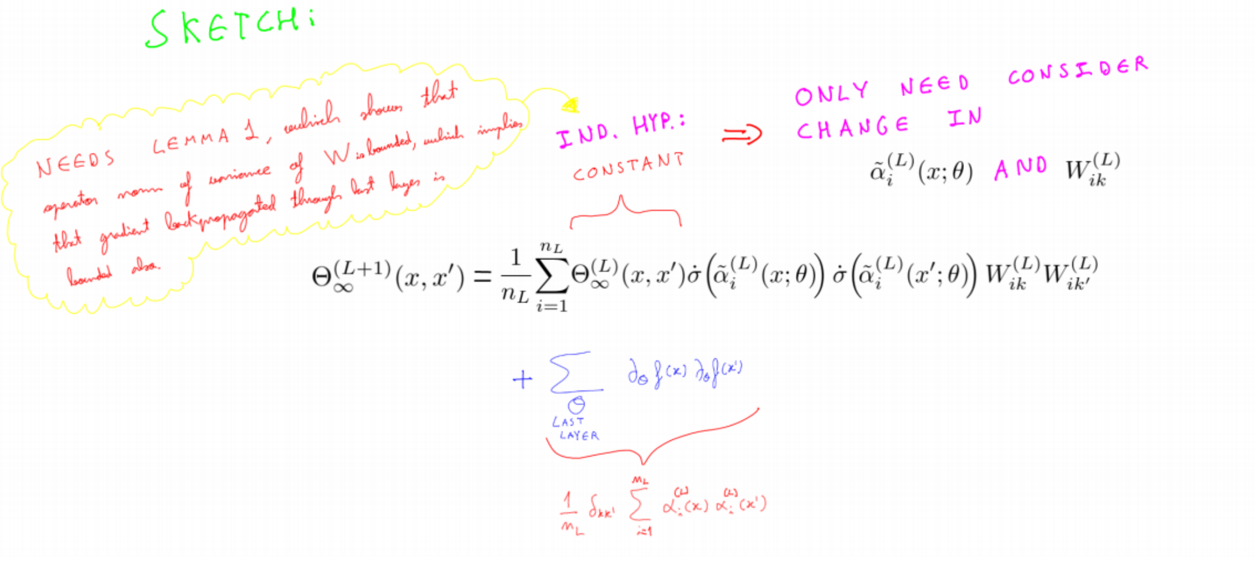
Theorem 2. *Assume that σ is a Lipschitz, twice differentiable nonlinearity function, with bounded second derivative. For any T such that the integral $\int_0^T \|d_t\|_{p^{in}} dt$ stays stochastically bounded, as $n_1, \ldots, n_{L-1} \to \infty$ sequentially, we have, uniformly for $t \in [0,T]$,*

$$\Theta^{(L)}(t) \to \Theta_\infty^{(L)} \otimes Id_{n_L}.$$

As a consequence, in this limit, the dynamics of f_θ is described by the differential equation

$$\partial_t f_{\theta(t)} = \Phi_{\Theta_\infty^{(L)} \otimes Id_{n_L}}\left(\langle d_t, \cdot \rangle_{p^{in}}\right).$$

We again prove this by induction. We give a proof sketch in the following. The starting point is the recursive equation to express $\Theta_\infty^{(L+1)}(x,x')$ in terms of $\Theta_\infty^{(L)}(x,x')$ (which we basically derived in previous section)



SKETCH:

NEEDS LEMMA 1, which shows that operator norm of variance of W is bounded, which implies that gradient backpropagated through last layer is bounded also.

IND. HYP.: CONSTANT $\Longrightarrow$ ONLY NEED CONSIDER CHANGE IN $\tilde{\alpha}_i^{(L)}(x;\theta)$ AND $W_{ik}^{(L)}$

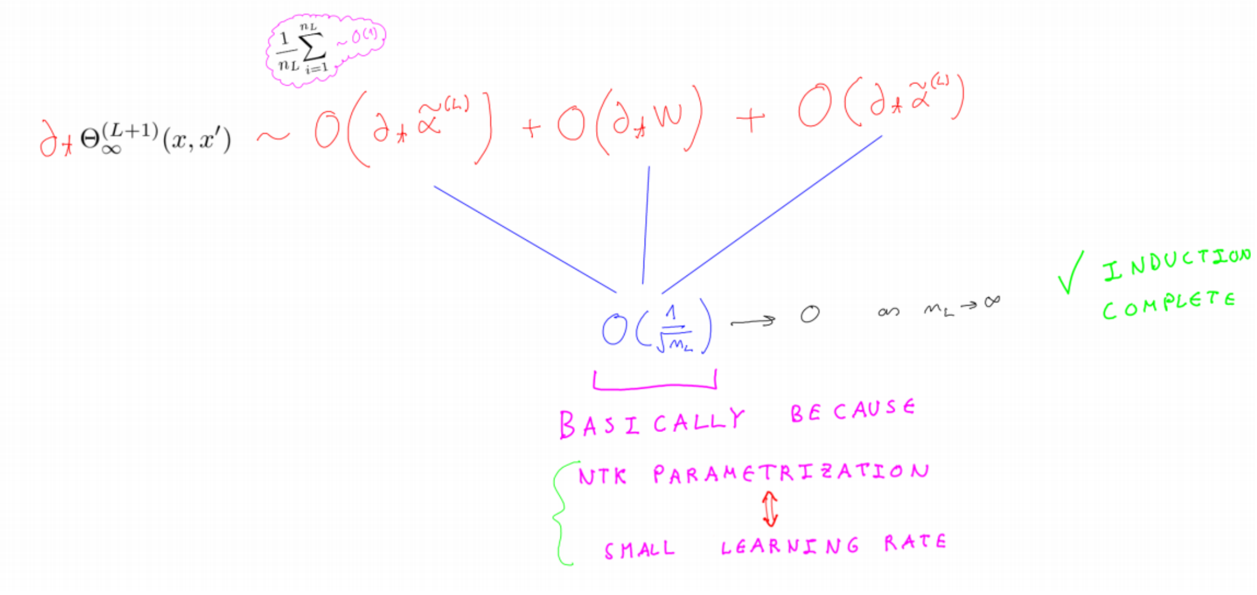
$$\Theta_\infty^{(L+1)}(x,x') = \frac{1}{n_L}\sum_{i=1}^{n_L}\overbrace{\Theta_\infty^{(L)}(x,x')}^{\text{IND. HYP.: CONSTANT}}\dot{\sigma}\left(\tilde{\alpha}_i^{(L)}(x;\theta)\right)\dot{\sigma}\left(\tilde{\alpha}_i^{(L)}(x';\theta)\right)W_{ik}^{(L)}W_{ik'}^{(L)}$$

$$+ \underbrace{\sum_{\substack{\theta \\ \text{LAST LAYER}}} \partial_\theta f(x)\, \partial_\theta f(x')}_{\frac{1}{n_L}\delta_{kk'}\sum_{i=1}^{n_L}\alpha_i^{(L)}(x)\,\alpha_i^{(L)}(x')}$$

That is: by the inductive hypothesis, the last layer NTK is constant[2], so we can write the **time derivative of the NTK** at the $L+1$ th layer as time derivatives of the last layer weights and the previous layer's preactivations alone.

The bulk of the proof (apart from the technical Lemma 1), is showin that the time derivatives w.r.t. to these weights and activations is $O(\frac{1}{\sqrt{n_L}})$. This is fundamentally because the NTK parametrization makes the weights evolve as if they have a very small learning rate, a feature we mentioned in the introduction and which we further discuss in Section 8

[2] this requires proving a lemma, Lemma 1, to show that the loss derivative doesn't blow up when back-propagated through the last layer


$\frac{1}{n_L}\sum_{i=1}^{n_L} \sim O(1)$
$\partial_t \Theta_\infty^{(L+1)}(x,x') \sim O\left(\partial_t \tilde{\alpha}^{(L)}\right) + O\left(\partial_t W\right) + O\left(\partial_t \tilde{\alpha}^{(L)}\right)$
$O\left(\frac{1}{\sqrt{n_L}}\right) \rightarrow 0$ as $n_L \rightarrow \infty$
√ INDUCTION COMPLETE
BASICALLY BECAUSE
NTK PARAMETRIZATION
⇕
SMALL LEARNING RATE

Remark. Lemma 1 bounds the deviation of the weigths W in operator (a.k.a. spectral) norm. This lemma is a weaker result than what is proven in the proof of the main theorem, which is a bound on the L^2 (aka Frobenius) norm.

The proof again proceeds by induction on L, the number of layers:

5.1 Base case

BASE CASE

When $L = 1$, the neural tangent kernel does not depend on the parameters, it is therefore constant during training.

5.2 Inductive step 1: Connecting gradient conditions

Let us remember the inductive hypothesis

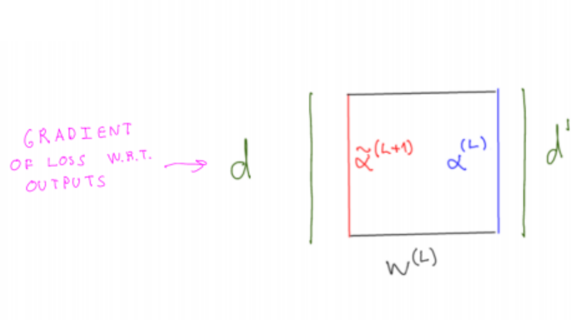
INDUCTION HYPOTHESIS

For any T such that the integral $\int_0^T \|d_t\|_{p^{in}} dt$ stays stochastically bounded as $n_1, ..., n_L \rightarrow \infty$ sequentially, we have, uniformly for $t \in [0,T]$,

$$\Theta^{(L)}(t) \rightarrow \Theta_\infty^{(L)} \otimes Id_{n_L}.$$

$$\partial_t f_{\theta(t)} = \Phi_{\Theta_\infty^{(L)} \otimes Id_{n_L}}\left(\langle d_t, \cdot \rangle_{p^{in}}\right).$$

The main condition of the inductive hypothesis requires that we bound the integral of d', the gradient of the loss with respect to the activations at the Lth layer. However, for induction, we only have the condition that the integral of d, the gradient of the loss w.r.t. activations at layer $L+1$, is bounded. We need to connect these two conditions, if we want to use the inductive hypothesis to prove the inductive step. Here is the equation of backpropagation



ELEMENT-WISE MULTIPLICATION

$$d'_t = \dot{\sigma}\left(\tilde{\alpha}^{(L)}(t)\right)\left(\frac{1}{\sqrt{n_L}}W^{(L)}(t)\right)^T d_t$$

And here is how we prove that the integral of d' is bounded from the fact that d is bounded. We use Cauchy-Schwartz, and then bound the norm of the weights. We bound the weights by bounding the norm of the weights at initialization, and then bounding the **operator norm** of the difference after a finite time $t < T$. Bounding the operator norm of the difference is done with Lemma 1, which we will prove later.

NEED $\int_0^T \|d'_t\|_{p^{in}} dt$ *stays stochastically bounded* ⊛

⇑

$\|d'_t\|_{p^{in}}$ *stays stochastically bounded*

CAUCHY-SCHWARTZ

$$\|d'_t\|_{p^{in}} \leq c\|\frac{1}{\sqrt{n_L}}W^{(L)}(t)\|_{op}\underbrace{\|d_t\|_{p^{in}}}_{\text{BOUNDED BY HYPOTHESIS IN THEOREM}}.$$

we now need to bound

$\|\frac{1}{\sqrt{n_L}}W^{(L)}(t)\|_{op}$

⇑

From this lemma

Lemma 1. *With the setting of Theorem 2, for a network of depth $L+1$, for any $\ell = 1, \ldots, L$, we have the convergence in probability:*

$$\lim_{n_L\to\infty}\cdots\lim_{n_1\to\infty}\sup_{t\in[0,T]}\|\frac{1}{\sqrt{n_\ell}}\left(W^{(\ell)}(t) - W^{(\ell)}(0)\right)\|_{op} = 0$$

⇑

it is hence enough to bound $\|\frac{1}{\sqrt{n_L}}W^{(L)}(0)\|_{op}$

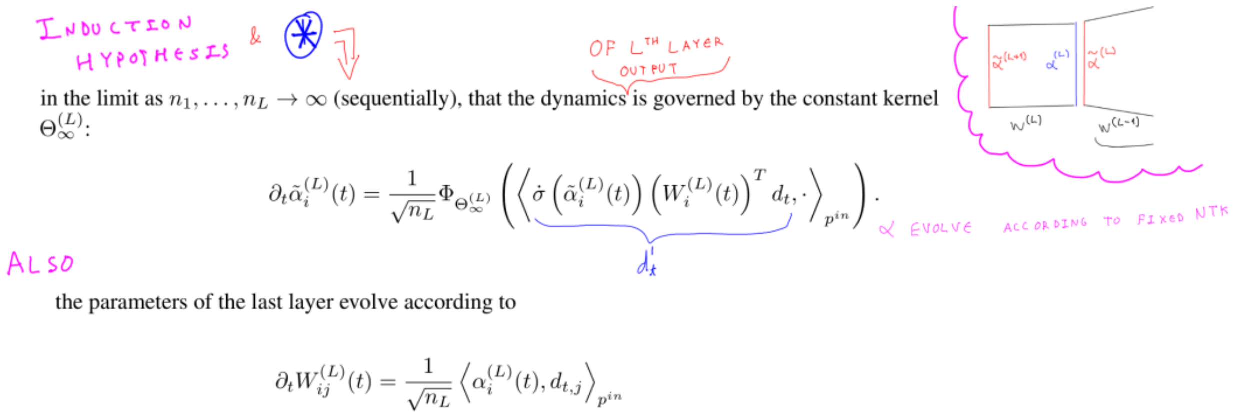
⇗

$$\|\frac{1}{\sqrt{n_L}}W^{(L)}(0)\|_{op} = \left\|\frac{1}{\sqrt{n_L}}\underbrace{\overbrace{\begin{pmatrix} \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \\ \cdot & \cdot & \cdots & \cdot & \cdot \end{pmatrix}}^{n_L}}_{\text{GAUSSIAN i.i.d.}} n_{L+1}\right\|_2 \overset{\text{law of large numbers}}{\xrightarrow[n_{L+1}\text{ fixed}]{n_L\to\infty}} \left\|\begin{pmatrix} O(1) \\ O(1) \\ O(1) \\ O(1) \end{pmatrix}\right\|_2 = O(1)$$

We have labelled the first condition by a star.

5.3 Inductive step 2: Last layer dynamics

To proceed we study the gradient flow dynamics of the previous-to-last layer activations, and the last layer weights.



INDUCTION HYPOTHESIS & ✱ ⇓

in the limit as $n_1, \ldots, n_L \to \infty$ (sequentially), that the dynamics is governed by the constant kernel $\Theta_\infty^{(L)}$:

OF L^{TH} LAYER OUTPUT

$$\partial_t \tilde{\alpha}_i^{(L)}(t) = \frac{1}{\sqrt{n_L}} \Phi_{\Theta_\infty^{(L)}} \left(\left\langle \underbrace{\dot{\sigma}\left(\tilde{\alpha}_i^{(L)}(t)\right)\left(W_i^{(L)}(t)\right)^T d_t}_{d_t'}, \cdot \right\rangle_{p^{in}} \right).$$

α EVOLVE ACCORDING TO FIXED NTK

ALSO

the parameters of the last layer evolve according to

$$\partial_t W_{ij}^{(L)}(t) = \frac{1}{\sqrt{n_L}} \left\langle \alpha_i^{(L)}(t), d_{t,j} \right\rangle_{p^{in}}$$

Here $\Phi_{\Theta_\infty^{(L)}(\langle d,\cdot\rangle_{p^{\text{in}}})(x)=\langle d,\Theta_\infty^{(L)(x,\cdot)}\rangle_{p^{\text{in}}}}$, as defined at the end of page 4 in the paper.

To show that NTK for layer $L+1$ becomes constant, we want to use these dynamics equations to show an upper bound on the variation of the last layer weights, in L^2 **norm**, and of the previous layer's activations (which are what appear in the recursive definition of the NTK of one layer given NTK of previous layer; see proof sketch above). Note that we want a bound in L^2 norm, rather than the operator norm which we have from Lemma 1, because we want to bound the variation in individual weights to show that the NTK doesn't change.

TO SHOW NTK FOR LAYER L+1 STAYS FIXED We want to USE $\partial_t\tilde{\alpha}_i^{(L)}(t) = \dots$ AND $\partial_t W_{ij}^{(L)}(t) = \dots$

to give an upper bound on the variation of the weights columns $W_i^{(L)}(t)$ and of the activations $\tilde{\alpha}_i^{(L)}(t)$ during training in terms of L^2-norm and p^{in}-norm respectively.

1) Cauchy-Schwarz inequality for each j,

$$\partial_t\left(W_{ij}^{(L)}(t) - W_i^{(L)}(0)\right) = \partial_t W_{ij}^{(L)}(t) \le \frac{1}{\sqrt{n_L}}||\alpha_i^{(L)}(t)||_{p^{in}}\,||d_{t,j}||_{p^{in}}$$

2) SQUARE AND SUM OVER J $||d_{t,j}||_{p^{in}} \to ||d_t||_{p^{in}}$.

3) and using $\partial_t||\cdot|| \le ||\partial_t \cdot||$

$$\rightarrow \quad \boxed{\partial_t\left\|W_i^{(L)}(t) - W_i^{(L)}(0)\right\|_2 \le \frac{1}{\sqrt{n_L}}||\alpha_i^{(L)}(t)||_{p^{in}}||d_t||_{p^{in}}}$$

The $\partial_t||\cdot|| \le ||\partial_t \cdot||$ can be obtained from the triangle inequality.

Now, observing that the operator norm of $\Phi_{\Theta_\infty^{(L)}}$ is equal to $||\Theta_\infty^{(L)}||_{op}$ defined in the introduction of Appendix A

and using the Cauchy-Schwarz inequality

where the sup norm $\|\cdot\|_\infty$ is defined by $\|f\|_\infty = \sup_x |f(x)|$.

$$\boxed{\partial_t\left\|\tilde{\alpha}_i^{(L)}(t) - \tilde{\alpha}_i^{(L)}(0)\right\|_{p^{in}} \le \frac{1}{\sqrt{n_L}}\left\|\Theta_\infty^{(L)}\right\|_{op}\underbrace{\left\|\dot{\sigma}\left(\tilde{\alpha}_i^{(L)}(t)\right)\right\|_\infty\left\|W_i^{(L)}(t)\right\|_2\|d_t\|_{p^{in}}}_{\begin{array}{c}\text{VI}\\ \left\|\dot{\sigma}\left(\tilde{\alpha}_i^{(L)}(t)\right)\left(W_i^{(L)}(t)\right)^T d_t\right\|_2\end{array}}}$$

These variations obey coupled differential inequalities. They are easier to analyze if we consider a single quantity which combines both, and which obeys a single differential inequality.

THE VARIATIONS, $\left\|W_i^{(L)}(t) - W_i^{(L)}(0)\right\|_2$ & $\left\|\tilde\alpha_i^{(L)}(t) - \tilde\alpha_i^{(L)}(0)\right\|_{p^{in}}$ OBEY COUPLE NONLINEAR ODEs (TRICKY)

CONSIDER SUM, BUT ∂_tS HAVE $||\alpha_i^{(L)}(t)||$ & $\left\|W_i^{(L)}(t)\right\|_2$,

SO WE ADD $||\alpha_i^{(L)}(0)||_{p^{in}}$ & $||W_i^{(L)}(0)||_2$ AND SEIZE TRIANGLE INEQ.

TO APPLY GRÖNWALL'S LEMMA

the derivative of the quantity $A(t) = ||\alpha_i^{(L)}(0)||_{p^{in}} + c\left\|\tilde\alpha_i^{(L)}(t) - \tilde\alpha_i^{(L)}(0)\right\|_{p^{in}} + ||W_i^{(L)}(0)||_2 + \left\|W_i^{(L)}(t) - W_i^{(L)}(0)\right\|_2.$

$$\partial_t A(t) \overset{|\dot\sigma| \le c}{\le} \frac{1}{\sqrt{n_L}}\left(c^2\left\|\Theta_\infty^{(L)}\right\|_{op}\left\|W_i^{(L)}(t)\right\|_2 + ||\alpha_i^{(L)}(t)||_{p^{in}}\right)||d_t||_{p^{in}} \overset{\triangle\text{ INEQ}}{\le} \frac{\max\{c^2\|\Theta_\infty^{(L)}\|_{op}, 1\}}{\sqrt{n_L}}\|d_t\|_{p^{in}}A(t),$$

We can then apply Grönwall's lemma to bound the variations

Applying Grönwall's Lemma

$$A(t) \le A(0)\exp\left(\overbrace{\frac{\max\{c^2\|\Theta_\infty^{(L)}\|_{op}, 1\}}{\sqrt{n_L}}\int_0^t \|d_s\|_{p^{in}} ds}^{\to 1}\right)$$

Note that $\|\Theta_\infty^{(L)}\|_{op}$ is constant during training. Clearly the value inside of the exponential converges to zero in probability as $n_L \to \infty$ given that the integral $\int_0^t \|d_t\|_{p^{in}} ds$ stays stochastically bounded.

$$\left\|\tilde\alpha_i^{(L)}(t) - \tilde\alpha_i^{(L)}(0)\right\|_{p^{in}} \leqslant c^{-1}(A(t) - A(0))$$
$$\left\|W_i^{(L)}(t) - W_i^{(L)}(0)\right\|_2 \leqslant A(t) - A(0)$$

$$\Rightarrow \boxed{\begin{aligned}\left\|\tilde\alpha_i^{(L)}(t) - \tilde\alpha_i^{(L)}(0)\right\|_{p^{in}} &\sim O\left(\tfrac{1}{\sqrt{n_L}}\right)\\ \left\|W_i^{(L)}(t) - W_i^{(L)}(0)\right\|_2 &\sim O\left(\tfrac{1}{\sqrt{n_L}}\right)\end{aligned}} \quad \text{as } n_L \to \infty$$

We can now use these bounds to control the variation of the NTK and to prove the theorem. We begin by studying the rate of change of the terms of the NTK corresponding to last layer's parameters

To understand how the NTK evolves, we study the evolution of the derivatives with respect to the parameters.

$$\left(\partial_{W^{(L)}_{ij}} f_{\theta,j'}(x) \otimes \partial_{W^{(L)}_{ij}} f_{\theta,j''}(x')\right)$$

The derivatives with respect to the bias parameters of the top layer $\partial_{b^{(L)}_j} f_{\theta,j'}$ are always equal to $\delta_{jj'}$

The derivatives with respect to the connection weights of the top layer are given by

$$\partial_{W^{(L)}_{ij}} f_{\theta,j'}(x) = \frac{1}{\sqrt{n_L}} \alpha^{(L)}_i(x;\theta)\delta_{jj'}. = O\left(\frac{1}{\sqrt{n_L}}\right)$$

The pre-activations $\tilde{\alpha}^{(L)}_i$ evolve at a rate of $\frac{1}{\sqrt{n_L}}$ and so do the activations $\alpha^{(L)}_i$. ∵ LIPSCHITZ

$$\partial_t \, \partial_{W^{(L)}_{ij}} f_{\theta,j'}(x) = \frac{1}{\sqrt{n_L}} \partial_t \alpha^{(L)}_i(x;\theta)\delta_{jj'} = O\left(\frac{1}{n_L}\right)$$

The summand $\partial_t\left(\partial_{W^{(L)}_{ij}} f_{\theta,j'}(x) \otimes \partial_{W^{(L)}_{ij}} f_{\theta,j''}(x')\right) = 2\,\partial_t\left(\partial_{W^{(L)}_{ij}} f_{\theta,j'}(x)\right) \otimes \partial_{W^{(L)}_{ij}} f_{\theta,j''}(x') = O\left(\frac{1}{n_L^{3/2}}\right)$

induce a variation of the NTK of rate $\frac{1}{\sqrt{n_L}}$.

$$\partial_t \Theta^{(L+1)}_{jj} \overset{\text{from } W^{(L)}}{=} \sum_{i=1}^{n_L} \partial_t\left(\partial_{W^{(L)}_{ij}} f_{\theta,j}(x) \otimes \partial_{W^{(L)}_{ij}} f_{\theta,j}(x')\right) = O\left(\frac{1}{\sqrt{n_L}}\right)$$

and now the rate of change of the terms of the NTK corresponding to lower layer's parameters

Finally let us study the derivatives with respect to the parameters of the lower layers

$$\partial_{\tilde{\theta}_k} f_{\theta,j}(x) = \frac{1}{\sqrt{n_L}} \sum_{i=1}^{n_L} \partial_{\tilde{\theta}_k} \tilde{\alpha}^{(L)}_i(x;\theta)\dot{\sigma}\left(\tilde{\alpha}^{(L)}_i(x;\theta)\right) W^{(L)}_{ij}.$$

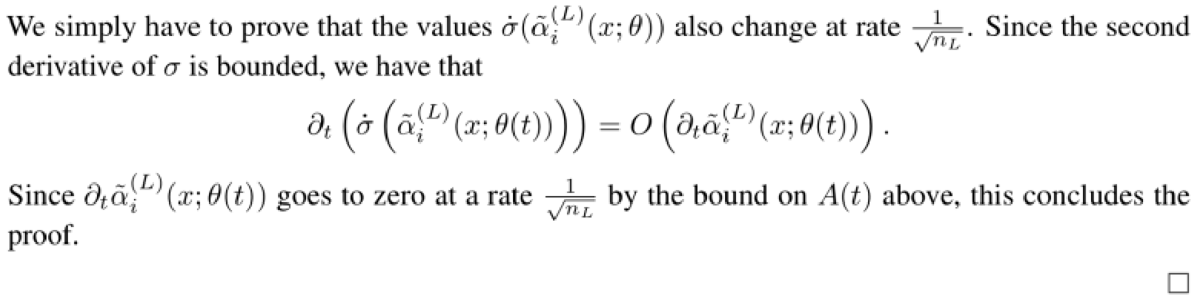
Their contribution to the NTK $\Theta^{(L+1)}_{jj'}(x,x')$ is

$$\frac{1}{n_L} \sum_{i,i'=1}^{n_L} \Theta^{(L)}_{ii'}(x,x')\dot{\sigma}\left(\tilde{\alpha}^{(L)}_i(x;\theta)\right)\dot{\sigma}\left(\tilde{\alpha}^{(L)}_{i'}(x';\theta)\right) W^{(L)}_{ij} W^{(L)}_{i'j'}.$$

By the induction hypothesis, the NTK of the smaller network $\Theta^{(L)}$ tends to $\Theta^{(L)}_\infty \delta_{ii'}$ as $n_1, ..., n_{L-1} \to \infty$. The contribution therefore becomes

$$\frac{1}{n_L} \sum_{i=1}^{n_L} \Theta^{(L)}_\infty(x,x')\dot{\sigma}\left(\tilde{\alpha}^{(L)}_i(x;\theta)\right)\dot{\sigma}\left(\tilde{\alpha}^{(L)}_i(x';\theta)\right) W^{(L)}_{ij} W^{(L)}_{ij'}.$$

The connection weights $W^{(L)}_{ij}$ vary at rate $\frac{1}{\sqrt{n_L}}$, inducing a change of the same rate to the whole sum.



We simply have to prove that the values $\dot{\sigma}(\tilde{\alpha}_i^{(L)}(x;\theta))$ also change at rate $\frac{1}{\sqrt{n_L}}$. Since the second derivative of σ is bounded, we have that

$$\partial_t \left(\dot{\sigma} \left(\tilde{\alpha}_i^{(L)}(x;\theta(t)) \right) \right) = O \left(\partial_t \tilde{\alpha}_i^{(L)}(x;\theta(t)) \right).$$

Since $\partial_t \tilde{\alpha}_i^{(L)}(x;\theta(t))$ goes to zero at a rate $\frac{1}{\sqrt{n_L}}$ by the bound on $A(t)$ above, this concludes the proof.

□

6 Lemma 1: Bounding the operator norm of weight variations

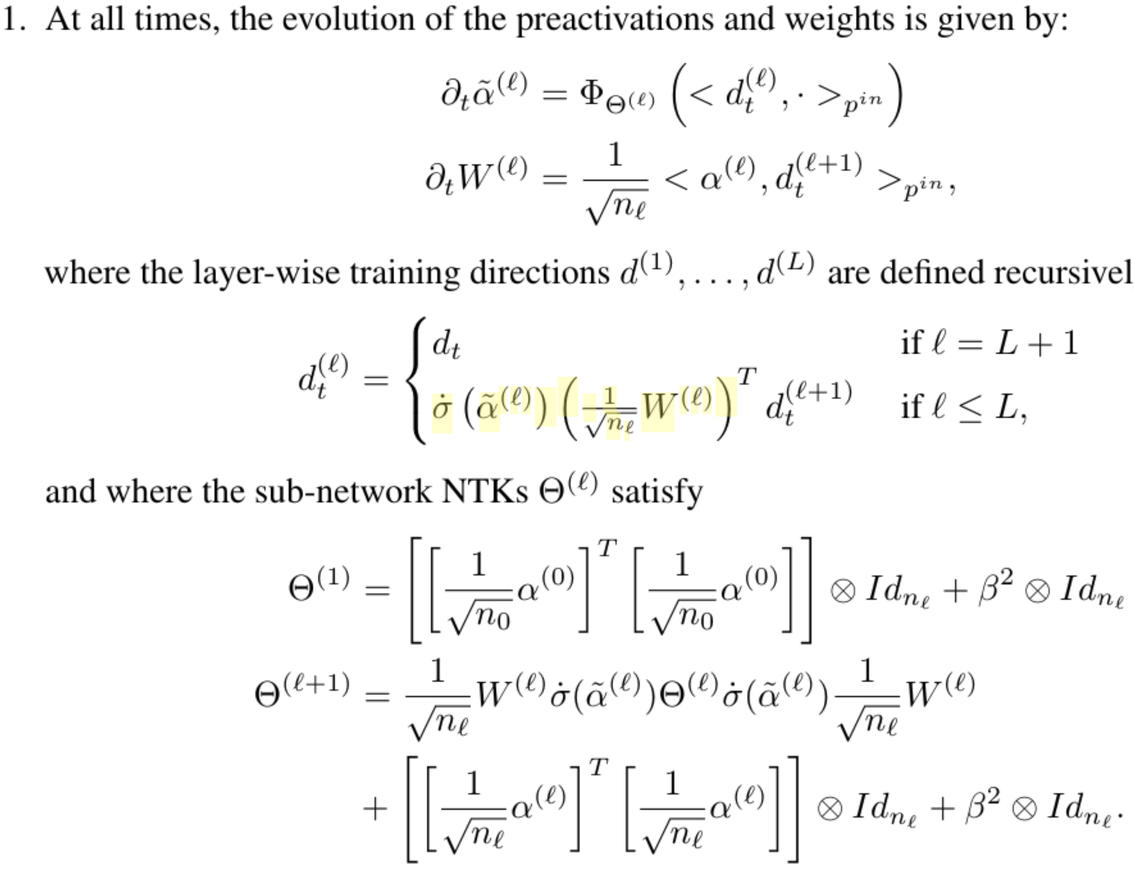
This lemma upper bounds the deviation in the weights in operator norm during training. This is useful to bound the L^2 norm of the backproagated gradients.

Lemma 1. With the setting of Theorem 2 for a network of depth $L+1$, for any $\ell = 1, \ldots, L, we$ have the convergence in probability:

$$\lim_{n_L \to \infty} \cdots \lim_{n_1 \to \infty} \sup_{t \in [0,T]} \left\| \frac{1}{\sqrt{n_\ell}} \left(W^{(\ell)}(t) - W^{(\ell)}(0) \right) \right\|_{op} = 0$$

We prove this, not by induction, but simultaneously for all $l = 1, \ldots, L$.

We first write the recursive equations for the backpropagated gradients, as well as the gradient flow equtations of the activations and weights, in terms of these backpropagated gradients. We also write the recursive equation for the NTK, which we have used sevreral times in the previous proofs; but now we write it in tensor notation, rather than with explicit sums over components



1. At all times, the evolution of the preactivations and weights is given by:

$$\partial_t \tilde{\alpha}^{(\ell)} = \Phi_{\Theta^{(\ell)}} \left(< d_t^{(\ell)}, \cdot >_{p^{in}} \right)$$
$$\partial_t W^{(\ell)} = \frac{1}{\sqrt{n_\ell}} < \alpha^{(\ell)}, d_t^{(\ell+1)} >_{p^{in}},$$

where the layer-wise training directions $d^{(1)}, \dots, d^{(L)}$ are defined recursively by

$$d_t^{(\ell)} = \begin{cases} d_t & \text{if } \ell = L+1 \\ \dot{\sigma}\left(\tilde{\alpha}^{(\ell)}\right) \left(\frac{1}{\sqrt{n_\ell}} W^{(\ell)} \right)^T d_t^{(\ell+1)} & \text{if } \ell \leq L, \end{cases}$$

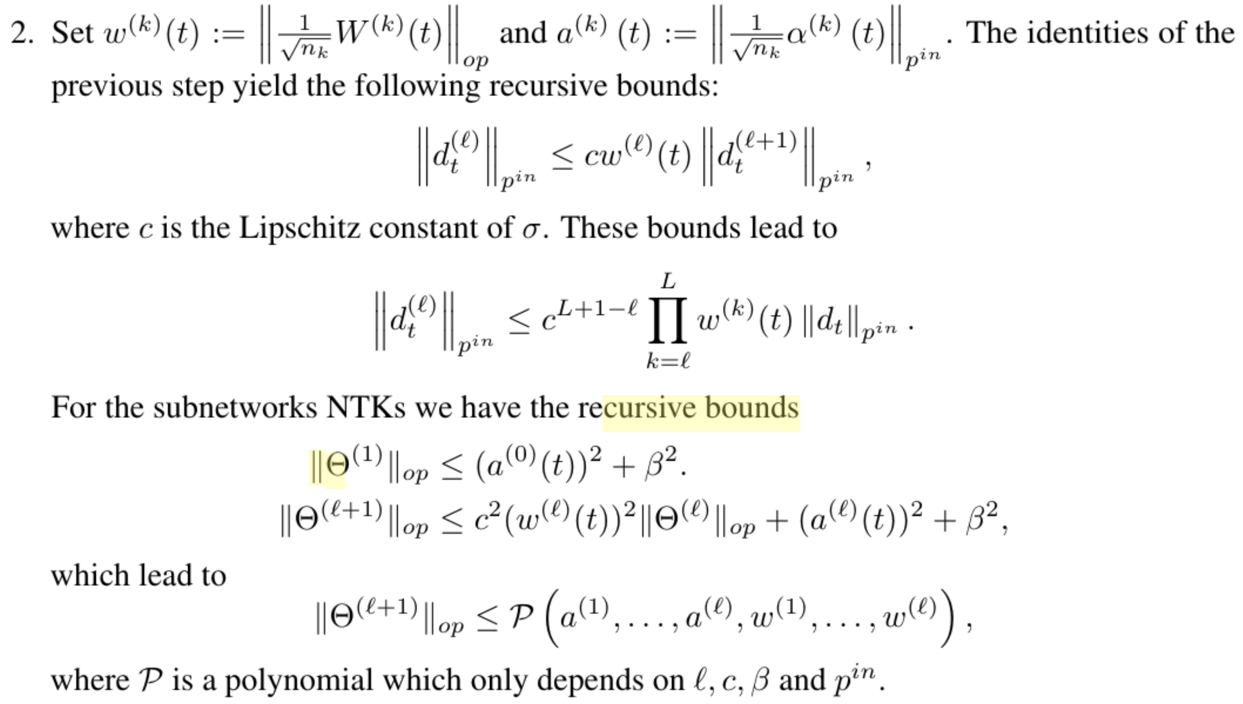
and where the sub-network NTKs $\Theta^{(\ell)}$ satisfy

$$\Theta^{(1)} = \left[\left[\frac{1}{\sqrt{n_0}} \alpha^{(0)} \right]^T \left[\frac{1}{\sqrt{n_0}} \alpha^{(0)} \right] \right] \otimes Id_{n_\ell} + \beta^2 \otimes Id_{n_\ell}$$
$$\Theta^{(\ell+1)} = \frac{1}{\sqrt{n_\ell}} W^{(\ell)} \dot{\sigma}(\tilde{\alpha}^{(\ell)}) \Theta^{(\ell)} \dot{\sigma}(\tilde{\alpha}^{(\ell)}) \frac{1}{\sqrt{n_\ell}} W^{(\ell)}$$
$$+ \left[\left[\frac{1}{\sqrt{n_\ell}} \alpha^{(\ell)} \right]^T \left[\frac{1}{\sqrt{n_\ell}} \alpha^{(\ell)} \right] \right] \otimes Id_{n_\ell} + \beta^2 \otimes Id_{n_\ell}.$$

The gradient flow eqution for $W^{(l)}$ I think is using a generalized definition of the inner product $\langle \cdot, \cdot \rangle_{p^{\text{in}}}$. In the intriduction this inner product is defined for pairs of functions $\mathbb{R}^{n_0} \to \mathbb{R}^{n_l}$. However, here the two functions have different codomains, $\mathbb{R}^{n_l}$ and $\mathbb{R}^{n_{l+1}}$. To make the equation true, I think one needs to define this as $\langle f, g \rangle_{p^{\text{in}}} := \mathbb{E}_{x \sim p^{\text{in}}}[f(x) g^T(x)]$, that is where we take the outer product of f and g rather than the inner product in the definition at the end of page 2.

Note that the identities for the $\Theta^{(1)}$ term above should be Id_{n_0}, not Id_{n_l}.

These recursive equations give us bounds on the norm of gradients, and on the operator norm of the NTK in terms of a polynomial of the norms of the activations and the weights. We use the definition of the operator norm to obtain the



2. Set $w^{(k)}(t) := \left\|\frac{1}{\sqrt{n_k}}W^{(k)}(t)\right\|_{op}$ and $a^{(k)}(t) := \left\|\frac{1}{\sqrt{n_k}}\alpha^{(k)}(t)\right\|_{p^{in}}$. The identities of the previous step yield the following recursive bounds:

$$\left\|d_t^{(\ell)}\right\|_{p^{in}} \leq cw^{(\ell)}(t)\left\|d_t^{(\ell+1)}\right\|_{p^{in}},$$

where c is the Lipschitz constant of σ. These bounds lead to

$$\left\|d_t^{(\ell)}\right\|_{p^{in}} \leq c^{L+1-\ell}\prod_{k=\ell}^{L} w^{(k)}(t)\left\|d_t\right\|_{p^{in}}.$$

For the subnetworks NTKs we have the recursive bounds

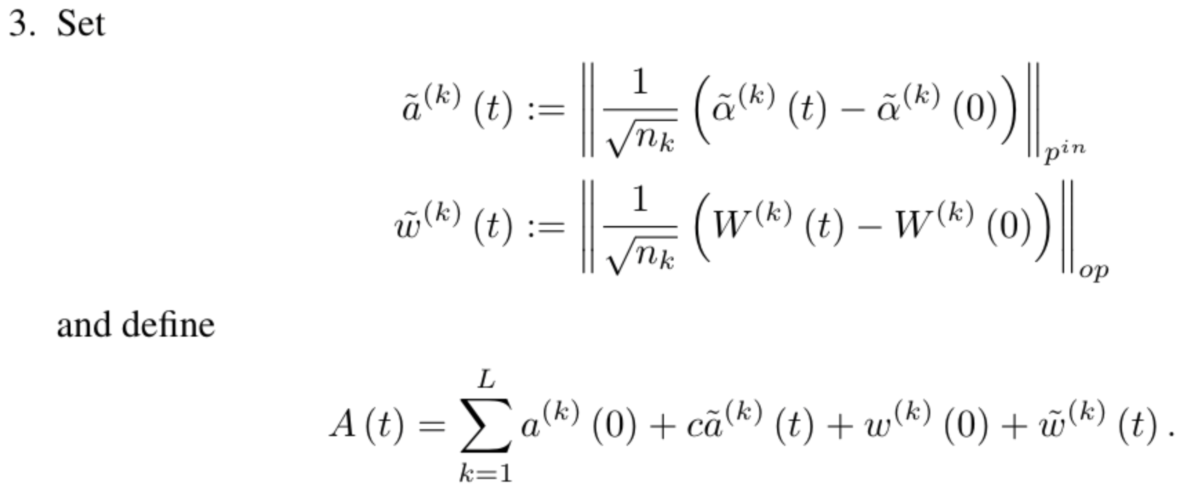
$$\|\Theta^{(1)}\|_{op} \leq (a^{(0)}(t))^2 + \beta^2.$$
$$\|\Theta^{(\ell+1)}\|_{op} \leq c^2(w^{(\ell)}(t))^2\|\Theta^{(\ell)}\|_{op} + (a^{(\ell)}(t))^2 + \beta^2,$$

which lead to

$$\|\Theta^{(\ell+1)}\|_{op} \leq \mathcal{P}\left(a^{(1)},\ldots,a^{(\ell)},w^{(1)},\ldots,w^{(\ell)}\right),$$

where $\mathcal{P}$ is a polynomial which only depends on ℓ, c, β and p^{in}.

We now define the relevant norms of the variations, as well as a quantity that combines them, that will allow us to bound them



3. Set

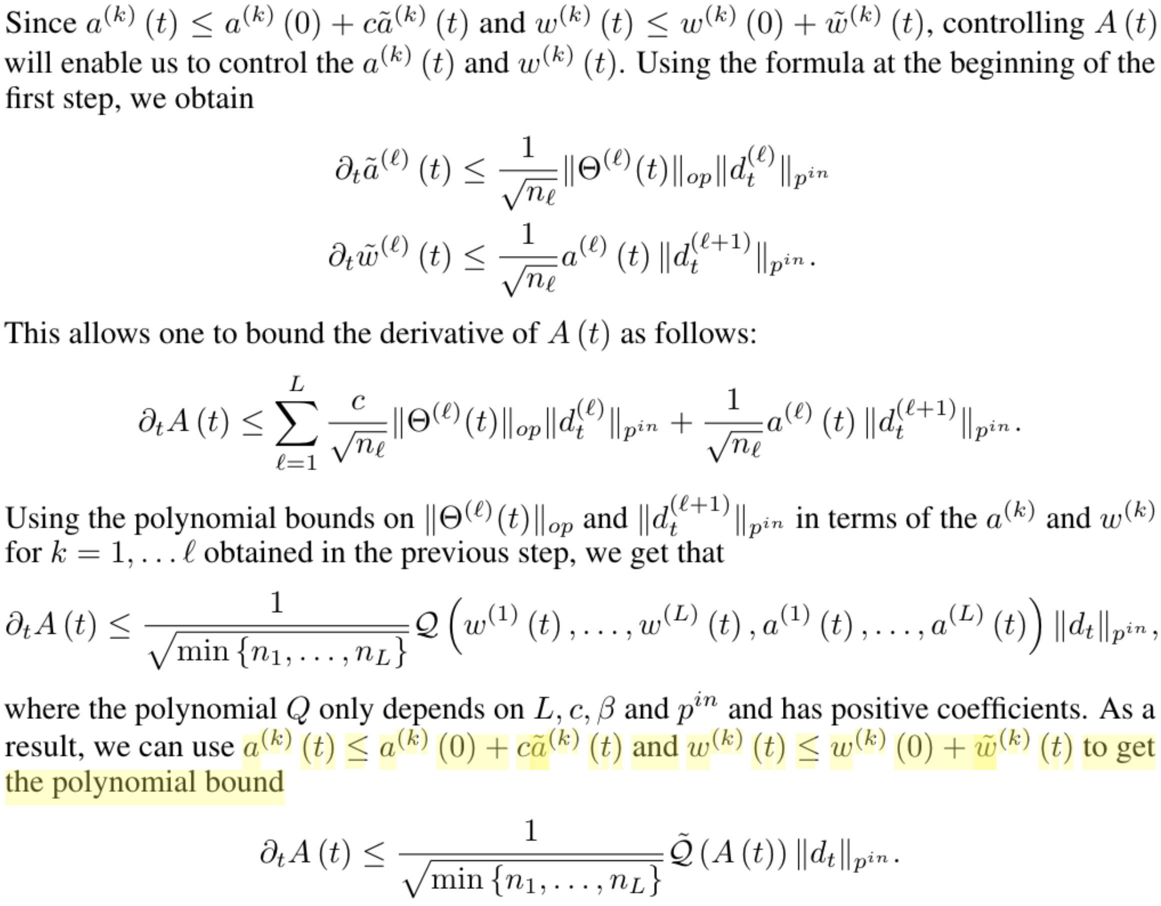
$$\tilde{a}^{(k)}(t) := \left\|\frac{1}{\sqrt{n_k}}\left(\tilde{\alpha}^{(k)}(t) - \tilde{\alpha}^{(k)}(0)\right)\right\|_{p^{in}}$$
$$\tilde{w}^{(k)}(t) := \left\|\frac{1}{\sqrt{n_k}}\left(W^{(k)}(t) - W^{(k)}(0)\right)\right\|_{op}$$

and define

$$A(t) = \sum_{k=1}^{L} a^{(k)}(0) + c\tilde{a}^{(k)}(t) + w^{(k)}(0) + \tilde{w}^{(k)}(t).$$

Finally, we convert the gradient flow equations into differential inequalities of the norms. Here we have used Cauchy-Schwartz to relate $\tilde{w}^{(l)}$ to $||W^{(l)}||_{op}$, we then use $\partial_t||\cdot|| \leq ||\partial_t \cdot||$, and finally we bound the operator norm of the right hand side, which is (omitting extra fluff) $\langle\alpha, d\rangle_{p^{\text{in}}} := \sum_x \alpha(x)d^T(x)$

The spectral norm of this (by considering its definition as maximum of $x^T Ax/||x||^2$) is bounded by $\sum_x ||\alpha(x)||||d(x)|| \leq \left(\sum_x ||\alpha(x)||^2\right)\left(\sum_x ||d(x)||^2\right) = ||\alpha||_{p^{\text{in}}}||d||_{p^{\text{in}}}$. This same idea is used in proving the bounds above.



Since $a^{(k)}(t) \leq a^{(k)}(0) + c\tilde{a}^{(k)}(t)$ and $w^{(k)}(t) \leq w^{(k)}(0) + \tilde{w}^{(k)}(t)$, controlling $A(t)$ will enable us to control the $a^{(k)}(t)$ and $w^{(k)}(t)$. Using the formula at the beginning of the first step, we obtain

$$\partial_t \tilde{a}^{(\ell)}(t) \leq \frac{1}{\sqrt{n_\ell}} \|\Theta^{(\ell)}(t)\|_{op} \|d_t^{(\ell)}\|_{p^{in}}$$

$$\partial_t \tilde{w}^{(\ell)}(t) \leq \frac{1}{\sqrt{n_\ell}} a^{(\ell)}(t) \|d_t^{(\ell+1)}\|_{p^{in}}.$$

This allows one to bound the derivative of $A(t)$ as follows:

$$\partial_t A(t) \leq \sum_{\ell=1}^{L} \frac{c}{\sqrt{n_\ell}} \|\Theta^{(\ell)}(t)\|_{op} \|d_t^{(\ell)}\|_{p^{in}} + \frac{1}{\sqrt{n_\ell}} a^{(\ell)}(t) \|d_t^{(\ell+1)}\|_{p^{in}}.$$

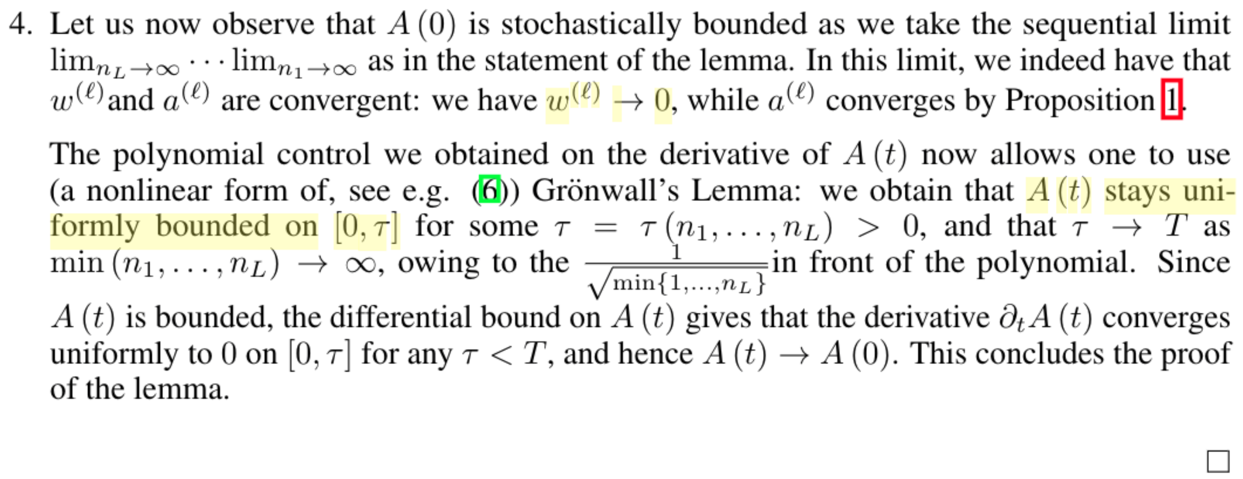
Using the polynomial bounds on $\|\Theta^{(\ell)}(t)\|_{op}$ and $\|d_t^{(\ell+1)}\|_{p^{in}}$ in terms of the $a^{(k)}$ and $w^{(k)}$ for $k = 1, \ldots \ell$ obtained in the previous step, we get that

$$\partial_t A(t) \leq \frac{1}{\sqrt{\min\{n_1, \ldots, n_L\}}} \mathcal{Q}\left(w^{(1)}(t), \ldots, w^{(L)}(t), a^{(1)}(t), \ldots, a^{(L)}(t)\right) \|d_t\|_{p^{in}},$$

where the polynomial Q only depends on L, c, β and p^{in} and has positive coefficients. As a result, we can use $a^{(k)}(t) \leq a^{(k)}(0) + c\tilde{a}^{(k)}(t)$ and $w^{(k)}(t) \leq w^{(k)}(0) + \tilde{w}^{(k)}(t)$ to get the polynomial bound

$$\partial_t A(t) \leq \frac{1}{\sqrt{\min\{n_1, \ldots, n_L\}}} \tilde{\mathcal{Q}}(A(t)) \|d_t\|_{p^{in}}.$$

I'm not sure what form of $\tilde{\mathcal{Q}}$ they had in mind, but substituting every a and w with $A(t)$ seems to work, although it would be quite a loose bound... Finally, we apply a nonlinear version of Grönwall's lemma to argue that the derivative converges uniformly to 0 on $fort \in 0, T, andhencethat$A(t)→ $A(0)$ and thus that the variations on a and w converge to 0.



4. Let us now observe that $A(0)$ is stochastically bounded as we take the sequential limit $\lim_{n_L\to\infty}\cdots\lim_{n_1\to\infty}$ as in the statement of the lemma. In this limit, we indeed have that $w^{(\ell)}$and $a^{(\ell)}$ are convergent: we have $w^{(\ell)} \to 0$, while $a^{(\ell)}$ converges by Proposition 1.

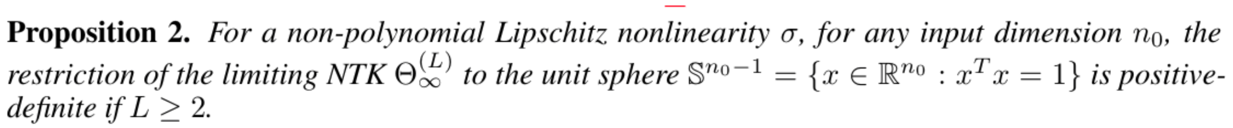
   The polynomial control we obtained on the derivative of $A(t)$ now allows one to use (a nonlinear form of, see e.g. (6)) Grönwall's Lemma: we obtain that $A(t)$ stays uniformly bounded on $[0,\tau]$ for some $\tau = \tau(n_1,\dots,n_L) > 0$, and that $\tau \to T$ as $\min(n_1,\dots,n_L) \to \infty$, owing to the $\frac{1}{\sqrt{\min\{1,\dots,n_L\}}}$in front of the polynomial. Since $A(t)$ is bounded, the differential bound on $A(t)$ gives that the derivative $\partial_t A(t)$ converges uniformly to 0 on $[0,\tau]$ for any $\tau < T$, and hence $A(t) \to A(0)$. This concludes the proof of the lemma.

□

I think the way to see that $w \to 0$ at initialization as claimed, we need to perform an analysis like that in the proof of Lemma 1 in [4]. We see that the operator norm is $O(\sqrt{n})$ for any $n \times m$ matrix. Therefore taking the limit $m \to \infty$ for the matrix multiplied by $1/\sqrt{m}$ before taking the limit $n \to \infty$, makes the operator norm converge to 0.

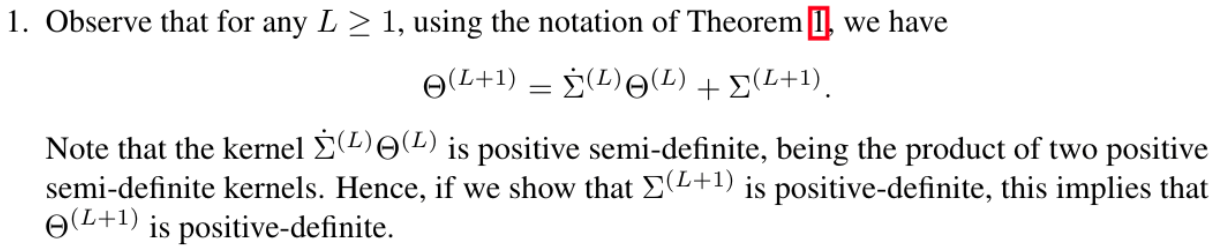
7 Positive-definiteness of $\Theta_\infty^{(L)}$

Jacot et al. also show that the NTK is positive definite when the inputs are restricted to the unit sphere.



Proposition 2. *For a non-polynomial Lipschitz nonlinearity σ, for any input dimension n_0, the restriction of the limiting NTK $\Theta_\infty^{(L)}$ to the unit sphere $\mathbb{S}^{n_0-1} = \{x \in \mathbb{R}^{n_0} : x^T x = 1\}$ is positive-definite if $L \geq 2$.*

The main idea of the proof is to reduce it to showing the positive definitness of the NTK at the second layer, as follows



1. Observe that for any $L \geq 1$, using the notation of Theorem 1, we have

$$\Theta^{(L+1)} = \dot{\Sigma}^{(L)}\Theta^{(L)} + \Sigma^{(L+1)}.$$

   Note that the kernel $\dot{\Sigma}^{(L)}\Theta^{(L)}$ is positive semi-definite, being the product of two positive semi-definite kernels. Hence, if we show that $\Sigma^{(L+1)}$ is positive-definite, this implies that $\Theta^{(L+1)}$ is positive-definite.

2. By definition, with the notation of Proposition 1 we have

$$\Sigma^{(L+1)}(x,x') = \mathbb{E}_{f\sim\mathcal{N}\left(0,\Sigma^{(L)}\right)}\left[\sigma(f(x))\,\sigma(f(x'))\right] + \beta^2.$$

This gives, for any collection of coefficients $c_1,\ldots,c_d \in \mathbb{R}$ and any pairwise distinct $x_1,\ldots,x_d \in \mathbb{R}^{n_0}$, that

$$\sum_{i,j=1}^{d} c_i c_j \Sigma^{(L+1)}(x_i,x_j) = \mathbb{E}\left[\left(\sum_i c_i\sigma(f(x_i))\right)^2\right] + \left(\beta\sum_i c_i\right)^2.$$

Hence the left-hand side only vanishes if $\sum c_i\sigma(f(x_i))$ is almost surely zero. If $\Sigma^{(L)}$ is positive-definite, the Gaussian $(f(x_i))_{i=1,\ldots d}$ is non-degenerate, so this only occurs when $c_1 = \cdots = c_d = 0$ since σ is assumed to be non-constant. This shows that the positive-definiteness of $\Sigma^{(L+1)}$ is implied by that of $\Sigma^{(L)}$. By induction, if $\Sigma^{(2)}$ is positive-definite, we obtain that all $\Sigma^{(L)}$ with $L \geq 2$ are positive-definite as well. By the first step this hence implies that $\Theta^{(L)}$ is positive-definite as well.

The the use a couple of technical results: a theorem on the postivie definitness of kernels with infinitely many positive terms in its power series expansion, which can be obtained by applyig a lemma about “duals of Lipschitz functions”.

8 Intuition on NTK results

Let us, for simplicity, consider a neural network with a single output, so that the function the network implements is $f : \mathbb{R}^{n_0} \to \mathbb{R}$.

What change in f does a small change in the parameters make? Let's Taylor expand the change in f on powers of the change in parameters $\Delta\theta$, to find out.

$\Delta f = J\Delta\theta + \Delta\theta^T H\Delta\theta + O(\Delta\theta^3)$

where J is the Jacobian, H is the Hessian.

Before analyzing how these behave as we make the network wider, let's think of a simpler situation studied in [1]. They consider what happens as you rescale the model output $f \mapsto af$, for some $a > 0$, and we simultaneously make the learning rate smaller with rate $O(1/a)$. In that case the Jacobian term stays of order $O(1)$, while the Hessian is of order $O(1/a)$, and goes to zero as $a \to \infty$. If furthermore, the Jacobian is of maximum rank, this means that Δf can take values over a non-degenerate ellipsoid of radius $O(1)$ even though $\Delta\theta$ is infinitesimally small. This means that we can fit any vector of target values that is within the $O(1)$ ellipsoid around the initial f[3], by changing the weights by an amount small enough that f is still approximately equal to its first order in the

[3] As long as the target vector is separated by an $O(1)$ distance from the initial f, we can always rescale the output f by a further constant to be able to enlarge the ellipsoid of possible values of f for fixed norm of $\Delta\theta$

Taylor expansion, that is, its linearization. If the learning rate is sufficiently small, this is what gradient flow (and in fact gradient descent too) will do.

But we aren't rescaling the output in this case. We are making the layers wider. We want to figure out the scaling of the Jacobian, and higher order terms, with the layer width, n. Let us begin with the Jacobian. Using the backpropagation equation, we can write the Jacobian as a product of matrices, with element-wise multiplications between every product, corresponding to the derivatives of the activations. These element-wise factors are $O(1)$, and so we can ignore them to get the order (the resulting expression would in fact be exact for linear activations). The partial derivative of the output w.r.t. to the ijth weight in layer $l < L$ (out of L layers) is

$$\frac{\partial f}{\partial W^{(l)}_{ij}} \sim \frac{1}{n^{\frac{L-l+1}{2}}} W^{(L)}_{k_L} W^{(L-1)}_{k_L k_{L-1}} \cdots W^{(l+1)}_{k_{l+2} i} \alpha^{(l)}_j \tag{2}$$

$$\sim \frac{1}{n^{L/2}} W^{(L)}_{k_L} W^{(L-1)}_{k_L k_{L-1}} \cdots W^{(l+1)}_{k_{l+2} i} \frac{1}{\sqrt{n}} W^{(l-1)}_{j k_{l-1}} W^{(l-2)}_{k_{l-1} k_{l-2}} \cdots W^{(0)}_{k_1 k_0} x_{k_0} \tag{3}$$

$$= O(\frac{1}{n}) \tag{4}$$

(remember we are using Einstein's summation convention). The factors of $1/\sqrt{n}$ came because we are using NTK parametrization. Now, we can use the same arguments as in section 3 to argue that each term in the sums should be i.i.d., and apply the central limit theorem to every sum over repeated indices, which should give us something of order $O(\sqrt{n})$, canceling out one of the factors of $1/\sqrt{n}$. However, because the derivative removed one layer's weight, there are two factors left, resulting in the $O(1/n)$. The result is different for the last layer weights, however, which have derivative of $O(1/\sqrt{n})$ because their contributions aren't "dampened" by a next layer of small weights. This agrees with the order of magnitude of the time derivatives of last layer weights used in the proof in section 5

If we apply the same argument to the Hessian elements $\frac{\partial^2 f}{\partial W^{(l')}_{i'j'} \partial W^{(l)}_{ij}}$, we would get $O(1/n^2)$. for most elements of the Hessian. However, the analysis is complicated because the element-wise factors we omitted above depend on the weights. The contribution from these can be argued[4] to also be $O(1/n^2)$. Again, when one of the weights are in the last layer, we'd get $O(1/n^{3/2})$, and 0 if both weights are in the last layer.

We now consider the product $J\Delta\theta$, because the gradient descent of each parameter $W^{(l)}_{ij}$ is proportional to $\frac{\partial f}{\partial W^{(l)}_{ij}}$, then

$$\begin{aligned} J\Delta\theta &\propto \sum_\theta \left(\frac{\partial f}{\partial \theta}\right)^2 \\ &= \sum_{\tilde{\theta}} O(\frac{1}{n^2}) + \sum_{\hat{\theta}} O(\frac{1}{n}) \\ &= O(1), \end{aligned}$$

[4] just consider the derivative with respect to an activation and then the derivative of the activation w.r.t. to the second weight. This is $O(1/n^{5/2})$. Summing over that layer's activations derivatives, and assuming these are also i.i.d., will give a contribution $O(1/n^2)$ per layer above the layer of the weight being differentiated

where $\hat{\theta}$ refers to the last layer's weights, and we used that there are $O(n^2)$ weights in the lower $L-1$ layers, and $O(n)$ weights in the last layer. Note that each term in the sum is positive so the sum increases the order by one power of n.

For the $\Delta\theta^T H \Delta\theta$ term, if we assume that its elements are independent of the elements of $\Delta\theta$, we can perform a simiar analysis.

$$\begin{aligned}\Delta\theta^T H \Delta\theta &\propto \sum_{\tilde{\theta}}\sum_{\tilde{\theta}} O(\frac{1}{n^2})O(\frac{1}{n^2}) + \sum_{\tilde{\theta}}\sum_{\hat{\theta}} O(\frac{1}{n^{3/2}})O(\frac{1}{n^{3/2}}) \\ &= O(\frac{1}{n^{3/2}}),\end{aligned}$$

as there are $O(n^4)$ and $O(n^3)$ terms in the first and second sums, respectively, and we square root these powers to obtain the contribution from the sum assuming i.i.d. terms. It seems that the second sum dominates. Nonetheless, the result goes to 0 as $n \to \infty$. I expect a similar analysis to carry thrhough for higher order terms, so that the first order term indeed dominates in the infinite width limit, using NTK parametrization. The same arguments that we presented above for rescaling the output can be applied now, to explain linear behavior during training to convergence. I also think the above arguments could be made more rigorous. I have seen maybe a couple of papers looking at the Hessian in the mean field (infinite width) limit. I wonder if they are related to my arguments here. I also think the above arguments could be made more rigorous. I have seen maybe a couple of papers looking at the Hessian in the mean field (infinite width) limit. I wonder if they are related to my arguments here. I also think the above arguments could be made more rigorous. I have seen maybe a couple of papers looking at the Hessian in the mean field (infinite width) limit. I wonder if they are related to my arguments here. I also think the above arguments could be made more rigorous. I have seen maybe a couple of papers looking at the Hessian in the mean field (infinite width) limit. I wonder if they are related to my arguments here.

9 NTK solution for least squares regression

Comming soon..

References

[1] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. 2019.

[2] Arthur Jacot, Franck Gabriel, and Clément Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In *Advances in neural information processing systems*, pages 8571–8580, 2018.

[3] Jaehoon Lee, Lechao Xiao, Samuel S Schoenholz, Yasaman Bahri, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *arXiv preprint arXiv:1902.06720*, 2019.

[4] Terry Tao. 254a, notes 3: The operator norm of a random matrix.