Final Paper Review

Brandon McKinzie

SUNet ID: 06009508

Abstract

A review of Lee et al. [2019] and Simon et al. [2021].

1. Problem Statement

Recent research in deep learning has aimed to formalize why heavily overparameterized deep neural networks seem to learn functions that generalize well to unseen data. Both Lee et al. [2019] and Simon et al. [2021] provide characterizations for training dynamics and generalization of deep neural networks that help explain such observed phenomena for overparameterized networks. Specifically, if we could derive a bound on the discrepancy between the deep networks used in practice vs linearized versions that are more amenable to analysis, we may be able to formalize the generalization properties of such overparameterized networks. This is the approach taken by Lee et al. [2019].

In Simon et al. [2021], the authors build upon recent work related to the Neural Tangent Kernel (NTK) and analyze the eigensystem of the NTK within the framework of kernel regression. Specifically, for a given network architecture, a target function f, and a training set of n random examples, can we efficiently predict the generalization performance of a network's learned function $\hat{f}$? More than just explaining why certain neural network architectures generalize well, understanding this would allow one to characterize which function classes a given architecture is well-suited to learn.

2. Main Result

Note that for each subsequent section, I provide a corresponding portion of the appendix containing the full derivations with more commentary.

2.1 Main Result of Lee et al. [2019]

Here, we'll focus on theorem 2.1 of Lee et al. [2019], restated below as Theorem 2.1, which says that with MSE loss, and when trained with a learning rate η below some critical learning rate η_{critical}, in the infinite-width limit the outputs of a neural network converge to its linearized counterpart. We will let $f_t(\mathcal{X}) \in \mathbb{R}^{k|\mathcal{X}|}$ denote the (vectorized) network's k outputs at gradient descent step t over all points $x \in \mathbb{R}^{n_0}$ in the training set $\mathcal{X}$, and let f_t^{lin} denote the outputs of its linearized counterpart, defined as

$$f_t^{lin}(x) \triangleq f_0(x) + \nabla_\theta f_0(x)\big|_{\theta=\theta_0}(\theta_t - \theta_0) \tag{1}$$

We assume that f is a feedforward network parameterized by θ.

Theorem 2.1 *Let $n_1 = \cdots = n_L = n$, where n_i denotes the width of layer i, and assume $\lambda_{min}(\Theta) > 0$, where Θ is the neural tangent kernel (NTK) (see 61). Applying gradient descent with learning rate $\eta < \eta_{critical}$ (or gradient flow) results in the following: $\forall x \in \mathbb{R}^{n_0}$ such that $||x||_2 \leq 1$, with probability arbitrarily close to 1 over random initialization,*

$$\sup_{t\geq 0} \left|\left| f_t(x) - f_t^{lin}(x) \right|\right|_2 = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) \tag{2}$$

$$\sup_{t\geq 0} \frac{1}{\sqrt{n}} \left|\left| \theta_t - \theta_0 \right|\right|_2 = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) \tag{3}$$

$$\sup_{t\geq 0} \left|\left| \hat{\Theta}_t - \hat{\Theta}_0 \right|\right|_F = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) \tag{4}$$

as $n \to \infty$.

Informally, this theorem can be summarized as three statements:

1. As $n \to \infty$, the outputs of a network f_t at any step t in gradient descent on an arbitrary point x converge to the outputs of its linearized counterpart $f_t^{lin}(x)$.
2. As $n \to \infty$, the parameters of the network at step t converge to the parameters at initialization.
3. As $n \to \infty$, the empirical NTK at step t converges to the empirical NTK at initialization[1].

The derivation hinges on first proving 3. Although it's difficult to think in the realm of the infinite, we can roughly interpret 3 as being a consequence of the fact that sums of the form $\sum_i^n \theta_i h_i^{(\ell)}$ can vary dramatically for even infinitesimal changes in θ when the summation is over infinitely many terms.

Given statement (2) above (corresponding to equation 3), the other two statements are perhaps less surprising; if the weights don't diverge significantly from their initial values over the course of training, one would intuitively think that the gradients of the network would remain fairly close to the gradients at initialization (equation 4). Similarly, it makes sense that if the gradients aren't fluctuating much compared to their values at initialization, then the higher order derivatives of f_t (on average over the course of training) should be similarly small, meaning that f_t is well approximated by f_t^{lin} in such cases, too (equation 2).

2.2 Main Result of Simon et al. [2021]

If we accept that infinite-width networks are linearized networks, it is natural to wonder what types of functions such networks learn. In Simon et al. [2021], the authors analyze the learned function of a infinite-width network trained via gradient descent to zero training MSE loss:

$$\hat{f}(x) = \Theta(x, \mathcal{D})\Theta(\mathcal{D}, \mathcal{D})^{-1} f(\mathcal{D}) \tag{5}$$

where f denotes the underlying target function we want to learn and $\mathcal{D}$ denotes a training set of n elements, and we assume a discretized input space with total cardinality M. The paper hinges on a quantity they define as the **learnability** of arbitrary target function f.

$$\mathcal{L}^{(\mathcal{D})}(f) \triangleq \frac{\left\langle f, \hat{f} \right\rangle}{\langle f, f \rangle} \qquad \mathcal{L}(f) \triangleq \mathbb{E}_{\mathcal{D}}\left[\mathcal{L}^{(\mathcal{D})}(f)\right] \tag{6}$$

[1] Since the tangent kernel of a finite-width network depends on the specific random draw of the parameters θ, the authors refer to it as the *empirical* tangent kernel to distinguish it from the *analytic* NTK Θ.

where $\mathcal{L}^{(\mathcal{D})}$ is referred to as the dataset-dependent learnability, or *$\mathcal{D}$-learnability.* Aside from having many nice properties, such as $\mathcal{L}(f) \in [0, 1]$, with

$$n = 0 \implies \mathcal{L}(f) = 0 \tag{7}$$

$$n = M \implies \mathcal{L}(f) = 1 \tag{8}$$

the learnability is also an intuitive framework for understanding the limiting behavior of networks trained to learn a given target function f. The result that I'll focus on shows that, using a certain decomposition of $\mathcal{L}$ in the eigenbasis of the NTK,

$$\forall i \in \{1, \cdots, M\} \qquad \frac{d}{d\lambda_i}\mathcal{L}^{(\mathcal{D})}(\phi_i) \geq 0 \tag{9}$$

$$\forall i, j \neq i \in \{1, \cdots, M\} \qquad \frac{d}{d\lambda_i}\mathcal{L}^{(\mathcal{D})}(\phi_j) \leq 0 \tag{10}$$

which states that the eigenmodes are in a zero-sum competition with each other to be learned for a given dataset $\mathcal{D}$. Increasing the eigenvalue λ_i for a given eigenfunction ϕ_i *improves* the learnability for that eigenfunction[2], but *harms* the learnability of all other eigenfunctions. As we'll see, this interpretation hinges on the main theorem from the paper, which the authors call the *"No-free-lunch" theorem for kernel regression*:

Theorem 2.2 ("No-free-lunch" theorem for kernel regression) *For any complete basis of orthogonal functions $\{\phi_i\}$*

$$\sum_i \mathcal{L}(\phi_i) = \sum_i \mathcal{L}^{(\mathcal{D}_n)}(\phi_i) = n \tag{11}$$

The authors are quick to emphasize that this is a much stronger theorem than previous results, which required averaging over *all* target functions f rather than an arbitrary set of basis functions $\{\phi_i\}$. This constraint, that the sum total learnability over the eigenbasis is always exactly n, is responsible for the zero-sum competition amongst eigenmodes.

3. Proof Sketch of Theorem 2.1

Here I provide a highly condensed proof sketch of 2.1 (from Lee et al. [2019]). In section B of the appendix, I provide a complete proof derivation that's reorganized for clarity and includes intermediate steps/interpretations not provided by the authors' derivation. The proof consists of two main steps:

1. Prove the global convergence of overparameterized neural networks and stability of the NTK under gradient descent.
2. Couple the stability of the NTK with Grönwall's type arguments to upper bound $\sup_{t\geq 0} \left|\left| f_t(x) - f_t^{lin}(x)\right|\right|_2$.

In the infinite-width limit, the outputs of a neural network can be expressed as a Gaussian Process Neal [1996]. Accordingly, the outputs $f(\mathcal{X}, \theta_0)$ on the training points $\mathcal{X}$ at initialization is itself jointly Gaussian. The authors use this to derive a bound of the form[3],

$$||f(\mathcal{X}, \theta_0) - \mathcal{Y}||_2 < R_0 \qquad (\text{w.p. } 1 - \delta_0) \tag{12}$$

[2] Previous works have shown that neural networks exhibit a spectral bias to eigenmodes with larger eigenvalues Canatar et al. [2021]. In Simon et al. [2021], the authors further show that these eigenmodes are in a zero-sum competition with one another for the ability to be learned at a given training set size.

[3] See 63 in the appendix for a more complete statement of 12

where R_0 is a constant[4]. This, in conjunction with the local Lipschitzness of the Jacobian (Lemma B.1), allows us to perform a proof by induction on t to obtain bounds on $||\theta_{t+1} - \theta_t||_2$ for arbitrary t. A simple application of the triangle inequality can extend this to a bound on $||\theta_t - \theta_0||_2$, yielding equation 3.

Next, we aim to prove 4 which gives us the stability of the NTK. As an intermediate step, we'll need to obtain an upper bound for $||f(\mathcal{X}, \theta_t) - \mathcal{Y}||_2$, which can again be done with inductive arguments on t, using the Gaussian Process result for obtaining the base case bound $||f(\mathcal{X}, \theta_0) - \mathcal{Y}||_2$. The main step here involves a telescoping sum and invocation of the mean value theorem. Using the shorthand $g(\theta_{t+1}) \equiv f(\mathcal{X}, \theta_{t+1}) - \mathcal{Y}$,

$$\begin{aligned} ||g(\theta_{t+1})||_2 &= ||g(\theta_{t+1}) - g(\theta_t) + g(\theta_t)||_2 && (13) \\ &= \left|\left|J(\tilde{\theta}_t)(\theta_{t+1} - \theta_t) + g(\theta_t)\right|\right|_2 \qquad \textbf{[MVT]} && (14) \\ &\leq \left|\left|1 - \eta J(\tilde{\theta}_t) J(\theta_t)^\top\right|\right|_{op} \left(1 - \frac{\eta_0 \lambda_{\min}}{3}\right)^t R_0 \qquad [57] && (15) \end{aligned}$$

If we can bound the operator norm of the last equation, we obtain the desired result. This is accomplished by exploiting the definition of Θ as converging to $\hat{\Theta}_0$ in the infinite-width limit, decomposing the norm above into telescoping sums of Θ and $\hat{\Theta}_0$, and again using the local Lipschitzness of the gradients. It is then straightforward to apply the same logic to bound $\left|\left|\hat{\Theta}_t - \hat{\Theta}_0\right|\right|_F$. The last and arguably most important part of the proof is deriving an upper bound on $\left|\left|f_t^{lin}(x) - f_t(x)\right|\right|_2$ for all t. Let $h(t) \triangleq \exp(\eta_0 \hat{\Theta}_0 t)\left(\hat{\Theta}_t - \hat{\Theta}_0\right)$.

$$\frac{d}{dt}\left(\exp(\eta_0 \hat{\Theta}_0 t)\left(g^{lin}(t) - g(t)\right)\right) = \eta_0 h(t) g(t) \tag{16}$$

$$g^{lin}(t) - g(t) = -\int_0^t \eta_0 h(s-t)\left(g^{lin}(s) - g(s)\right) \mathrm{d}s + \int_0^t \eta_0 h(s-t) g(s) \mathrm{d}s \tag{17}$$

Taking the norm of this, and denoting $\lambda_0 \equiv \lambda_{\min}(\hat{\Theta}_0)$,

$$e^{\eta_0 \lambda_0 t}\left|\left|g^{lin}(t) - g(t)\right|\right|_2 \leq \alpha(t) + \int_0^t \beta(s) u(s) \mathrm{d}s \leq \alpha(t) \exp\left(\int_0^t \beta(s) \mathrm{d}s\right) \tag{18}$$

for non-decreasing $\alpha(t)$. By leveraging the stability of the NTK derived earlier, we can obtain

$$\sup_t \left|\left|g^{lin}(t) - g(t)\right|\right|_2 \lesssim \sup_t \sigma_t R_0 \lesssim R_0{}^2 \frac{1}{\sqrt{n}} \tag{19}$$

where $\sigma_t \triangleq \sup_{0 \leq s \leq t}\left|\left|\hat{\Theta}_s - \hat{\Theta}_0\right|\right|_{op}$. Recall that $g(t) \equiv f(\mathcal{X}, \theta_t) - \mathcal{Y}$, i.e. so far we've only derived bounds for quantities over the *training* data $\mathcal{X}$. Our ultimate goal is obtaining such points for arbitrary inputs (re: test points) x. Luckily, due to the local Lipschitzness of the gradients (lemma B.1), we can state

$$\begin{aligned} \sup_t \left|\left|\hat{\Theta}_0(x, \mathcal{X}) - \hat{\Theta}_t(x, \mathcal{X})\right|\right|_2 &\lesssim R_0 \frac{1}{\sqrt{n}} && (20) \\ \implies \left|\left|g^{lin}(t, x) - g(t, x)\right|\right|_2 &\lesssim R_0 \frac{1}{\sqrt{n}} && (21) \end{aligned}$$

which completes the proof for 2 and thus the proof for theorem 2.1 (for the gradient flow case).

[4] That may depend on $\delta_0, |\mathcal{X}|, \mathcal{K}$, but that doesn't depend on the network width n.

Note that the paper does not provide the second part of the proof for the discrete gradient descent case. Below I sketch a proof for bounding the discrepancy in this case, and I'm able to get a bound of the same desired form ($\lesssim R_0{}^2 \frac{1}{\sqrt{n}}$). First, recall the basic formulas for f^{lin} and the change in outputs at step $t+1$ compared to step t:

$$f^{lin}(\theta_t) \triangleq f(\theta_t) + J(\theta_0)(\theta_t - \theta_0) \tag{22}$$

$$\implies f^{lin}(\theta_{t+1}) - f^{lin}(\theta_t) = J(\theta_0)(\theta_{t+1} - \theta_t) \tag{23}$$

$$||f_{t+1} - f_t||_2 \leq ||J(\theta_t)(\theta_{t+1} - \theta_t)||_2 \tag{24}$$

where 24 uses the triangle inequality to discard higher order terms.

$$\begin{aligned}
\left|\left|g^{lin}(t+1) - g(t+1) - (g^{lin}(t) - g(t))\right|\right|_2 &= ||(J(\theta_0) - J(\theta_t))(\theta_{t+1} - \theta_t)||_2 && (25)\\
&\leq ||J(\theta_0) - J(\theta_t)||_{op} ||\theta_{t+1} - \theta_t||_2 && (26)\\
&\leq K\sqrt{n}\, ||\theta_t - \theta_0||_2 \, ||\theta_{t+1} - \theta_t||_2 \qquad \textbf{[B.1]} && (27)\\
&\leq K\sqrt{n}\, ||\theta_t - \theta_0||_2 \frac{K\eta_0}{\sqrt{n}} \left(1 - \frac{\eta_0 \lambda_{\min}}{3}\right)^t R_0 && (28)\\
&\leq \eta_0 K^2 R_0 \left(1 - \frac{\eta_0 \lambda_{\min}}{3}\right)^t \frac{3KR_0}{\lambda_{\min}} \frac{1}{\sqrt{n}} && (29)\\
&= 3\frac{\eta_0 K^3 R_0{}^2}{\lambda_{\min}} \left(1 - \frac{\eta_0 \lambda_{\min}}{3}\right)^t \frac{1}{\sqrt{n}} && (30)
\end{aligned}$$

which gives us an $\mathcal{O}(\frac{1}{\sqrt{n}})$ bound on the *change* in discrepancy between gradient descent steps. We can combine this with the individual bounds derived previously on $\left|\left|g_t^{lin}\right|\right|_2$ and $||g_t||_2$, again using the triangle inequality and lemma B.1, to obtain the desired result.

4. Proof Sketch for Simon et al. [2021]

Here I'll provide a proof sketch for the result that the eigenmodes are in a zero-sum competition with each other in terms of learnability[5]. Formally, this means proving the property (g) of Lemma 1 from Simon et al. [2021]:

Lemma 4.1 *For any* $i, j \in \{1, \ldots, M\}, i \neq j$, $\frac{d}{d\lambda_i}\mathcal{L}^{(\mathcal{D})}(\phi_j) \leq 0$

Although the direct proof for this is short and simple, it requires sketching out the preliminary results derived earlier in the paper. Therefore, my proof sketch will essentially be showing how we arrive at this result when starting from the "beginning," which I'll define as equation 5. First, we need to express 5 using the notation of learnability. To do this, we convert to the eigenbasis of the NTK Θ. Let $\{\phi_i\}_{i=1}^M$ denote the M orthonormal eigenfunctions that satisfy

$$\frac{1}{M}\sum_{x' \in \mathcal{X}} \Theta(x, x')\phi_i(x') = \lambda_i \phi_i(x) \qquad \langle \phi_i, \phi_j \rangle = \delta_{i,j} \tag{31}$$

which follows from the Mercer decomposition of Θ, which I provide a brief review for in section D of the appendix. We can decompose the target function f and learned function $\hat{f}$ into this eigenbasis as

$$f(x) = \sum_{i=1}^M v_i \phi_i(x) \qquad \hat{f}(x) = \sum_{i=1}^M \hat{v}_i \phi_i(x) \tag{32}$$

[5] As before, I also include a much more complete derivation/discussion in the appendix (section C).

Henceforth, we'll use the vector notation $\boldsymbol{v}$ and $\hat{\boldsymbol{v}}$ in $\mathbb{R}^M$ for f and $\hat{f}$, respectively. Noting that Θ is diagonalized in this basis by definition, we can rewrite 5 in this vector space as

$$\hat{\boldsymbol{v}} \triangleq \underbrace{\boldsymbol{\Lambda}\boldsymbol{\Phi}(\mathcal{D}) \left(\boldsymbol{\Phi}^\top(\mathcal{D})\boldsymbol{\Lambda}\boldsymbol{\Phi}(\mathcal{D})\right)^{-1} \boldsymbol{\Phi}^\top(\mathcal{D})}_{\triangleq \boldsymbol{T}^{(\mathcal{D})}} \boldsymbol{v} \tag{33}$$

where $\boldsymbol{\Phi}_{ij}(\mathcal{D}) = \phi_i(x^{(j)})$ and $\boldsymbol{T}^{(\mathcal{D})} \in \mathbb{R}^{M \times M}$ is referred to as the *learning transfer matrix.* Note that $\boldsymbol{T}^{(\mathcal{D})}$ is independent of the target function f and thus fully captures the model's learning behavior on a training set $\mathcal{D}$. The key quantity in the paper, called the *$\mathcal{D}$-learnability* and denoted by $\mathcal{L}^{(\mathcal{D})}$ can be rewritten in this vector notation as

$$\mathcal{L}^{(\mathcal{D})}(f) \triangleq \frac{\left\langle f, \hat{f} \right\rangle}{\langle f, f \rangle} = \frac{\boldsymbol{v}^\top \hat{\boldsymbol{v}}}{|\boldsymbol{v}|^2} = \frac{\boldsymbol{v}^\top \boldsymbol{T}^{\mathcal{D}} \boldsymbol{v}}{|\boldsymbol{v}|^2} \tag{34}$$

Recall that our goal is to prove lemma 4.1. From equation 34, we see that

$$\mathcal{L}^{(\mathcal{D})}(\phi_i) \triangleq \frac{\left\langle \phi_i, \hat{\phi}_i \right\rangle}{\langle \phi_i, \phi_i \rangle} \tag{35}$$

$$= \boldsymbol{e}_i^\top \boldsymbol{T}^{(\mathcal{D})} \boldsymbol{e}_i \tag{36}$$

$$= \boldsymbol{T}_{ii} \tag{37}$$

$$\implies \frac{d}{d\lambda_i} \mathcal{L}^{(\mathcal{D})}(\phi_j) = \frac{d}{d\lambda_i} \boldsymbol{T}_{jj}^{(\mathcal{D})} \tag{38}$$

Therefore, since

$$\frac{d}{d\lambda_i} \boldsymbol{T}_{jj}^{(\mathcal{D})} = \left(\delta_{ij} - \lambda_j \phi_j^\top \Theta^{-1} \phi_i\right) \phi_i^\top \Theta^{-1} \phi_j \tag{39}$$

$$(i \neq j) \implies \frac{d}{d\lambda_i} \boldsymbol{T}_{jj}^{(\mathcal{D})} = -\lambda_j (\phi_j^\top \Theta^{-1} \phi_i)^2 \tag{40}$$

and since $\lambda_j > 0$ (from the assumption that Θ is positive definite), we have the desired result (lemma 4.1). It's worth taking a moment to recognize that, due to theorem 2.2, which can be equivalently stated as $\mathrm{Tr}(\boldsymbol{T}^{(\mathcal{D})}) = n$, an increase in the eigenvalue λ_i of ϕ_i simultaneously "causes" a decrease in the learnability of *all other eigenfunctions ϕ_j*.

5. Limitations

Perhaps the most obvious limitations of both Lee et al. [2019] and Simon et al. [2021] is (1) working in the infinite-width limit, and (2) focusing on MSE loss[6]. With that said, Simon et al. [2021] shows excellent agreement of the theory with finite width networks even for widths as small as 20. Beyond that, both papers include a large number of assumptions in the derivations of their proofs, and it can be challenging deciphering which assumptions are reasonable and broadly applicable vs assumptions that aren't well approximated in common real-world scenarios.

For example, let's interpret a few of the key assumptions mentioned by theorem 2.1. The theorem is restricted to inputs satisfying $||x||_2 = 1$. Upon closer inspection, this is only required insofar as allowing us to make statements regarding Lipschitzness of the gradients over the training set $\mathcal{X}$. Since the authors consider ReLU activation and linear readout layer, we couldn't bound much if $||x||_2$ wasn't itself bounded. Basically, as long as $||x||_2 \leq C$ for some universal constant C, the theorem still holds,

[6] To be fair, Lee et al. [2019] also includes some analysis of cross entropy in the appendix.

but it would be more cumbersome to include additional constants throughout the proof. Furthermore, normalizing the inputs to a network is commonly done in practice, and hence is a reasonable assumption.

Next, as noted by the authors, the restriction to $\{x \in \mathbb{R}^{n_0} : ||x||_2 = 1\}$ actually implies that $\lambda_{\min}(\Theta) > 0$[7], where Θ is the analytic NTK defined by

$$\Theta = \text{plim}_{n\to\infty} \frac{1}{n} \nabla_\theta f(\mathcal{X}, \theta_0) \nabla_\theta f(\mathcal{X}, \theta_0)^\top \tag{41}$$

To understand this, I considered the following decomposition of $\nabla_\theta f(\mathcal{X}, \theta_0) \nabla_\theta f(\mathcal{X}, \theta_0)^\top$ applied to one of the standard basis vectors $\{e_i\}_{i=1}^{k|\mathcal{X}|}$

$$\nabla_\theta f(\mathcal{X}, \theta_0) \nabla_\theta f(\mathcal{X}, \theta_0)^\top e_i = \sum_{p=1}^{|\theta|} \frac{\partial f(\mathcal{X}, \theta_0)}{\partial \theta_p} \frac{\partial f(\mathcal{X}, \theta_0)}{\partial \theta_p}^\top e_i \tag{42}$$

$$= \sum_{p=1}^{|\theta|} \frac{\partial f(\mathcal{X}, \theta_0)}{\partial \theta_p} \left(\frac{\partial f(\mathcal{X}, \theta_0)}{\partial \theta_p} \right)_i \tag{43}$$

Since we can express e_i in the eigenbasis of Θ, the constraint that $\lambda_{\min} > 0$ implies that for every output k on every training point $x \in \mathcal{X}$, denoted $f_0(x)_k$, there exists at least one parameter θ_p in the network for which $\frac{\partial f_0(x)_k}{\partial \theta_p} > 0$. At first glance, this might actually seem quite restrictive. However, we are working in the infinite-width limit, $|\theta| \gg k|\mathcal{X}|$, and thus requiring that the columns of $\nabla_\theta f(\mathcal{X}, \theta_0) \in \mathbb{R}^{k|\mathcal{X}| \times |\theta|}$ span $\mathbb{R}^{k|\mathcal{X}|}$ is quite reasonable.

6. Future directions

As noted elsewhere in the literature, one of the remaining puzzles that arises when comparing infinite-width predictions from theory with finite networks in practice is the seemingly contradictory observations that (1) we observe improving performance in finite-width networks as we increase their width, but (2) finite-width networks often outperform direct applications of their infinite-width counterparts. After all, if increasing the width of the network leads to improved generalization, and if we have analytic forms in the infinite-width limit, why isn't everyone strictly working with such networks in practice? For one, Lee et al. [2019] observed that the agreement between f and f^{lin} was much less stable when training with cross entropy loss compared to MSE. Understanding the underlying causes for discrepancies between theory and experiment in these settings is a promising direction for future research.

It's also worth mentioning that Simon et al. [2021] assumed a discrete input space of $|\mathcal{X}|$ unique points. In experiments on continuous input spaces, all but the top few eigenmodes had learnabilities much closer to 0 than 1, even when the number of data points was as high as 2^8 for the input space of the 7-sphere (Figure 2(c) of Simon et al. [2021]).This paints a slightly different perspective for both the limitations and future directions for studying the infinite width limit: the approximation is excellent when the target function is composed of the top eigenmodes of the NTK. Perhaps it is the case that, for architectures/tasks that aren't well suited for the infinite-width approximation, the target functions are poorly aligned with the top eigenmodes (i.e. give more weight to lower-eigenvalue modes). In any case, it would be useful for future experiments to analyze the relationship between kernel spectra and agreement between finite vs infinite-width networks, similar to what was done in Bahri et al. [2021] but with a focus on relating the decomposition of the target function in terms of the eigenmodes and the agreement with linearized approximations.

[7] Under the global assumptions of the paper that we are working with feedforward networks with activations $\phi(x)$ that grow non-polynomially in large x.

References

Yasaman Bahri, Ethan Dyer, Jared Kaplan, Jaehoon Lee, and Utkarsh Sharma. Explaining neural scaling laws. CoRR, abs/2102.06701, 2021. URL https://arxiv.org/abs/2102.06701.

Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Spectral bias and task-model alignment explain generalization in kernel regression and infinitely wide neural networks. Nature Communications, 12(1), May 2021. ISSN 2041-1723. doi: 10.1038/s41467-021-23103-1. URL http://dx.doi.org/10.1038/s41467-021-23103-1.

Jaehoon Lee, Lechao Xiao, Samuel S. Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent, 2019.

Radford M. Neal. Bayesian learning for neural networks. 1996. doi: https://doi.org/10.1007/978-1-4612-0745-0.

James B. Simon, Madeline Dickens, and Michael R. DeWeese. Neural tangent kernel eigenvalues accurately predict generalization, 2021.

A. Appendix

A.1 Review: Single-Layer MLPs as Gaussian Processes

A review of the results presented in Neal [1996]. The results here will be foundational for my complete proof sketch in the following section. Let $\boldsymbol{f_\theta} : \mathbb{R}^I \to \mathbb{R}^O$, where $\boldsymbol{\theta} = \{\boldsymbol{a}, \boldsymbol{U}, \boldsymbol{b}, \boldsymbol{V}\}$, be a neural network with one hidden layer of width H.

$$f_k(\boldsymbol{x}) = b_k + \sum_{j=1}^{H} V_{kj} h_j(\boldsymbol{x}) \tag{44}$$

$$h_j(\boldsymbol{x}) = \tanh\left(a_j + \sum_{i=1}^{I} U_{ji} x_i\right) \tag{45}$$

Assume $a_i \overset{i.i.d}{\sim} \mathcal{N}\left(0, \sigma_b^2\right)$, and $V_{k,j} \overset{i.i.d.}{\sim} \mathcal{N}\left(0, \sigma_v^2\right)$. For the moment, assume some fixed input $\boldsymbol{x} := \boldsymbol{x}^{(1)}$. Denote $\sigma_h^2 \triangleq \mathbb{E}\left[h_j(\boldsymbol{x}_1)^2\right]$, which is the same for all $1 \leq j \leq H$.

$$\begin{aligned}
\mathbb{E}\left[f_k(\boldsymbol{x}^{(1)})\right] &= \mathbb{E}\left[b_k\right] + \sum_{j=1}^{H} \mathbb{E}\left[V_{k,j} h_j(\boldsymbol{x}^{(1)})\right] &(46)\\
&= 0 + \sum_{j=1}^{H} 0 \cdot \mathbb{E}\left[h_j(\boldsymbol{x}^{(1)})\right] \qquad V_{j,k} \perp \{a_{j'}, U_{j',i}\} &(47)\\
&= 0 &(48)\\
\mathrm{Var}\left[f_k(\boldsymbol{x}^{(1)})\right] &= \sigma_b^2 + \sum_{j=1}^{H} \sigma_v^2 \mathrm{Var}\left[h_j(\boldsymbol{x}^{(1)})\right] &(49)\\
&= \sigma_b^2 + H\sigma_v^2\sigma_h^2 &(50)\\
&= \sigma_b^2 + \omega_v^2\sigma_h^2 \quad \text{where} \quad \sigma_v = \frac{1}{\sqrt{H}}\omega_v &(51)
\end{aligned}$$

Note that we have not made any assumptions about the distribution of the hidden layer weights/biases $\boldsymbol{V}$ and $\boldsymbol{a}$, aside from the fact that they are drawn element-wise-i.i.d. and are independent of the output layer weights/biases $\boldsymbol{V}$ and $\boldsymbol{b}$. By the Central Limit Theorem, we can say that as $H \to \infty$, $f_k(\boldsymbol{x}_1)$ converges to a random Gaussian variable with mean 0 and variance $\sigma_b^2 + \omega_v^2\sigma_h^2$.

Similarly, the joint prior of f_k over a set of inputs, $\{f_k(\boldsymbol{x}^{(1)}), \ldots, f_k(\boldsymbol{x}^{(n)})\}$ itself converges to a multivariate Gaussian with mean zero and covariances

$$\Sigma_{p,q} = \mathbb{E}\left[f_k(\boldsymbol{x}^{(p)}) f_k(\boldsymbol{x}^{(q)})\right] = \sigma_b^2 + \omega_v^2 C_{p,q} \tag{52}$$

$$\text{where } C_{p,q} = \mathbb{E}\left[h_j(\boldsymbol{x}^{(p)}) h_j(\boldsymbol{x}^{(q)})\right] \tag{53}$$

This property of f_k, that we can treat any [finite] collection of its values for a set of inputs $\{\boldsymbol{x}^{(i)}\}$, and that collection is itself jointly Gaussian, is the definition of a **Gaussian process**.

B. Complete Proof of Theorem 2.1

Here I present the full derivation of theorem 2.1, including far more detail than provided by Lee et al. [2019]. There were many gaps in the derivation (things the authors likely assumed the reader understood) that I had to re-derive/prove to myself, all of which are included below. First, I'll restate the key intermediate lemmas/theorems that are involved in the proof. An equation we'll reference often is the weight update for gradient descent:

$g(\theta_t) \in \mathbb{R}^{k|\mathcal{X}|}$

$$\theta_{t+1} - \theta_t = -\eta J(\theta_t)^\top g(\theta_t) \tag{54}$$

$J(\theta_t) \in \mathbb{R}^{k|\mathcal{X}|\times|\theta|}$

where $J(\theta_t) \triangleq \nabla_\theta f(\mathcal{X}, \theta_t)$.

Lemma B.1 (Local Lipschitzness of the Jacobian) *$\exists K > 0$ s.t. $\forall\, C > 0$, w.h.p. over random initialization (w.h.p.o.r.i), $\forall \theta, \tilde{\theta} \in B(\theta_0, \frac{1}{\sqrt{n}}\, C)$:*

$$\frac{1}{\sqrt{n}} \left|\left|J(\theta)\right|\right|_F \leq K \tag{55}$$

$$\frac{1}{\sqrt{n}} \left|\left|J(\theta) - J(\tilde{\theta})\right|\right|_F \leq K \left|\left|\theta - \tilde{\theta}\right|\right|_2 \tag{56}$$

Theorem B.2 (Gradient Descent) *For $(\ \delta_0 > 0)(\eta_0 < \eta_{critical})(\exists R_0 > 0, N \in \mathbb{N}, K > 1)$ such that $\forall n \geq N$, w.p. at least $1 -\ \delta_0$ over random initialization when applying GD with $\eta := \frac{\eta_0}{n}$:*

$$\left|\left|f(\theta_t) - \mathcal{Y}\right|\right|_2 \leq \left(1 - \frac{\eta_0 \lambda_{min}}{3}\right)^t R_0 \tag{57}$$

$$\sum_{j=1}^{t} \left|\left|\theta_j - \theta_{j-1}\right|\right|_2 \leq \frac{\eta_0 K R_0}{\sqrt{n}} \sum_{j=1}^{t} \left(1 - \frac{\eta_0 \lambda_{min}}{3}\right)^{j-1} \leq \frac{3 K R_0}{\lambda_{min}} \frac{1}{\sqrt{n}} \tag{58}$$

$$\sup_t \left|\left|\hat{\Theta}_0 - \hat{\Theta}_t\right|\right|_F \leq \frac{6K^3 R_0}{\lambda_{min}} \frac{1}{\sqrt{n}} \tag{59}$$

Theorem B.3 (Convergence to Linearized Network) *For all $x \in \mathbb{R}^{n_0}$ with $\left|\left|x\right|\right|_2 \leq 1$, for $\ \delta_0 > 0$ arbitrarily small, $\exists R_0 > 0$ and $N \in \mathbb{N}$ s.t. $\forall n \geq N$, w.p. at least $1 -\ \delta_0$ over random initialization,*

$$\sup_t \left|\left|g^{lin}(t) - g(t)\right|\right|_2,\ \sup_t \left|\left|g^{lin}(t, x) - g(t, x)\right|\right|_2 \lesssim {R_0}^2 \frac{1}{\sqrt{n}} \tag{60}$$

Assumptions

1. $n_1 = \cdots n_L = n$.
2. The analytic NTK Θ is full-rank.

$$\hat{\Theta}_t \equiv \hat{\Theta}_t(\mathcal{X}, \mathcal{X}) = \frac{1}{n} J(\theta_t) J(\theta_t)^\top \quad \text{and} \quad \Theta \triangleq \text{plim}_{n\to\infty} \hat{\Theta}_0 \tag{61}$$

3. $\mathcal{D}$ contained in some compact set and $x \neq \tilde{x}\ \forall x, \tilde{x} \in \mathcal{X}$.
4. The activation function ϕ satisfies

$$|\phi(0)|,\ \left|\left|\phi'\right|\right|_\infty,\ \sup_{x \neq \tilde{x}} \frac{|\phi'(x) - \phi'(\tilde{x})|}{|x - \tilde{x}|} < \infty \tag{62}$$

5. We are using MSE loss $\mathcal{L}(t) = \frac{1}{2}\left|\left|f(\mathcal{X}, \theta_t) - \mathcal{Y}\right|\right|_2^2$.

Bounds for Wide Networks (Outputs) at Initialization. Since $f(\mathcal{X}, \theta_t) \xrightarrow{p} \mathcal{N}(0, \mathcal{K})$ (Neal [1996]) one can show that for $\delta_0 > 0$, $\exists R_0(\delta_0, |\mathcal{X}|, \mathcal{K}) > 0$ and $\exists n_0(\delta_0, |\mathcal{X}|, \mathcal{K})$ s.t. $\forall n \geq n_0$, w.p. at least $1 - \delta_0$ over random initialization,

$$||f(\mathcal{X}, \theta_0) - \mathcal{Y}||_2 < R_0 \tag{63}$$

Per-Layer Parameter Convergence for Wide Networks. Prove 58 by induction: Choose $n_1 > n_0$ such that $\forall n \geq n_1$ equation 55, 56, and 63 hold w.p. at least $1 - \delta_0/5$. Assume 57 and 58 holds for some given t [8]. From equation 54, we know

$$\begin{aligned}
||\theta_{t+1} - \theta_t||_2 &\leq \eta \, ||J(\theta_t)||_{op} \, ||g(\theta_t)||_2 && (64)\\
&\leq \eta \, ||J(\theta_t)||_{op} \left(1 - \frac{\eta_0 \lambda_{\min}}{3}\right)^t R_0 \qquad \textbf{[57]} && (65)\\
&\leq \eta \, ||J(\theta_t)||_F \left(1 - \frac{\eta_0 \lambda_{\min}}{3}\right)^t R_0 \qquad [||A||_{op} \leq ||A||_F] && (66)\\
&\leq \frac{\eta_0}{n} K \sqrt{n} \left(1 - \frac{\eta_0 \lambda_{\min}}{3}\right)^t R_0 \qquad \textbf{[B.1]} && (67)\\
&= \frac{K \eta_0}{\sqrt{n}} \left(1 - \frac{\eta_0 \lambda_{\min}}{3}\right)^t R_0 && (68)\\
\implies ||\theta_{t+1} - \theta_0||_2 &\leq \sum_{j=1}^{t+1} ||\theta_j - \theta_{j-1}||_2 \qquad \textbf{[triangleq ineq.]} && (69)\\
&\leq \frac{\eta_0 K R_0}{\sqrt{n}} \sum_{j=1}^{t+1} \left(1 - \frac{\eta_0 \lambda_{\min}}{3}\right)^{j-1} \qquad \textbf{[58]} && (70)
\end{aligned}$$

Bounds for Wide Networks (Outputs) at Arbitrary Step. Prove 57 using the mean value theorem and 54. Again, proceed with proof by induction, assuming it holds for some given t, and show it holds for $t + 1$[9]:

$$\begin{aligned}
||g(\theta_{t+1})||_2 &= ||g(\theta_{t+1}) - g(\theta_t) + g(\theta_t)||_2 && (71)\\
&= \left|\left|J(\tilde{\theta}_t)(\theta_{t+1} - \theta_t) + g(\theta_t)\right|\right|_2 \qquad \textbf{[MVT]} && (72)\\
&= \left|\left|-\eta J(\tilde{\theta}_t) J(\theta_t)^\top g(\theta_t) + g(\theta_t)\right|\right|_2 \qquad \textbf{[54]} && (73)\\
&= \left|\left|(1 - \eta J(\tilde{\theta}_t) J(\theta_t)^\top) g(\theta_t)\right|\right|_2 && (74)\\
&\leq \left|\left|1 - \eta J(\tilde{\theta}_t) J(\theta_t)^\top\right|\right|_{op} ||g(\theta_t)||_2 \qquad [||Av||_2 \leq ||A||_{op} \, ||v||_2] && (75)\\
&\leq \left|\left|1 - \eta J(\tilde{\theta}_t) J(\theta_t)^\top\right|\right|_{op} \left(1 - \frac{\eta_0 \lambda_{\min}}{3}\right)^t R_0 \qquad \textbf{[57]} && (76)
\end{aligned}$$

Therefore, all that's left for us to prove 57 is to show that

$$\left|\left|1 - \eta J(\tilde{\theta}_t) J(\theta_t)^\top\right|\right|_{op} \leq 1 - \frac{\eta_0 \lambda_{\min}}{3} \tag{77}$$

[8] Note that they trivially satisfy the base induction case of $t := 0$

[9] Recall that we know it holds for $t = 0$ by 63.

Recall the definition of the analytic NTK:

$$\Theta \triangleq \operatorname{plim}_{n\to\infty} \hat{\Theta}_0 \tag{78}$$

It follows that $\exists n_2$ such that $\forall n \geq n_2$, w.p. at least $(1 - \delta_0/5)$,

$$\begin{aligned}
\left|\left|\Theta - \hat{\Theta}_0\right|\right|_F &\leq \frac{\eta_0 \lambda_{\min}}{3} && (79)\\
\implies ||1 - \eta_0 \Theta||_{op} &\leq 1 - \eta_0 \lambda_{\min} && (80)\\
\implies \left|\left|1 - \eta J(\tilde{\theta}_t) J(\theta_t)^\top\right|\right|_{op} &= \left|\left|1 - \eta J(\tilde{\theta}_t) J(\theta_t)^\top - \eta_0 \Theta + \eta_0 \Theta - \eta_0 \hat{\Theta}_0 + \eta_0 \hat{\Theta}_0\right|\right|_{op} && (81)\\
&\leq ||1 - \eta_0 \Theta||_{op} + \eta_0 \left|\left|\Theta - \hat{\Theta}_0\right|\right|_{op} + \left|\left|\eta_0 \hat{\Theta}_0 - \eta J(\tilde{\theta}_t) J(\theta_t)\right|\right|_{op} \qquad \textbf{[triangle ineq.]} && (82)\\
&\leq ||1 - \eta_0 \Theta||_{op} + \eta_0 \left|\left|\Theta - \hat{\Theta}_0\right|\right|_{op} + \left|\left|\frac{\eta_0}{n} J(\theta_0) J(\theta_0)^\top - \eta J(\tilde{\theta}_t) J(\theta_t)\right|\right|_{op} \qquad \textbf{[61]} && (83)\\
&\leq (1 - \eta_0)\, \lambda_{\min} + \frac{\eta_0 \lambda_{\min}}{3} + \eta \left|\left|J(\theta_0) J(\theta_0)^\top - J(\tilde{\theta}_t) J(\theta_t)\right|\right|_{op} && (84)\\
&\leq (1 - \eta_0)\, \lambda_{\min} + \frac{\eta_0 \lambda_{\min}}{3} + \eta_0 K^2 \left(||\theta_t - \theta_0||_2 + \left|\left|\tilde{\theta}_t - \theta_0\right|\right|_2\right) \qquad \textbf{[B.1]} && (85)\\
&\leq (1 - \eta_0)\, \lambda_{\min} + \frac{\eta_0 \lambda_{\min}}{3} + \eta_0 K^2 2 \frac{3 K R_0}{\lambda_{\min}} \frac{1}{\sqrt{n}} \qquad \textbf{[58]} && (86)\\
&= 1 - \frac{2}{3} \eta_0 \lambda_{\min} + 6 \frac{\eta_0 K^3 R_0}{\lambda_{\min} \sqrt{n}} && (87)
\end{aligned}$$

Recall that $\eta := \frac{\eta_0}{n}$

Remember, our goal was to show

$$\left|\left|1 - \eta J(\tilde{\theta}_t) J(\theta_t)^\top\right|\right|_{op} \leq 1 - \frac{\eta_0 \lambda_{\min}}{3} \tag{88}$$

and we have just shown that

$$\left|\left|1 - \eta J(\tilde{\theta}_t) J(\theta_t)^\top\right|\right|_{op} \leq 1 - \frac{2}{3} \eta_0 \lambda_{\min} + 6 \frac{\eta_0 K^3 R_0}{\lambda_{\min} \sqrt{n}} \tag{89}$$

Therefore, we can solve the following inequality for n

$$\begin{aligned}
1 - \frac{2}{3} \eta_0 \lambda_{\min} + 6 \frac{\eta_0 K^3 R_0}{\lambda_{\min} \sqrt{n}} &\leq 1 - \frac{\eta_0 \lambda_{\min}}{3} && (90)\\
\implies n &\geq \left(\frac{18 K^3 R_0}{\lambda_{\min}^2}\right)^2 && (91)
\end{aligned}$$

and we have thus proven 57 for

$$N := \max \left\{ n_0, n_1, n_2, \left(\frac{18 K^3 R_0}{\lambda_{\min}^2}\right)^2 \right\} \tag{92}$$

Stability of NTK under Gradient Descent. All that's left is to prove 59, after which we will have proven all of B.2, which was the first of two steps for proving our ultimate goal: theorem 2.1.

$$\left|\left|\hat{\Theta}_0 - \hat{\Theta}_t\right|\right|_F = \frac{1}{n}\left|\left|J(\theta_0)J(\theta_0)^\top - J(\theta_t)J(\theta_t)^\top\right|\right|_F \qquad \text{[61]} \tag{93}$$

$$= \frac{1}{n}\left|\left|J(\theta_0)J(\theta_0)^\top - J(\theta_0)J(\theta_t)^\top + J(\theta_0)J(\theta_t)^\top - J(\theta_t)J(\theta_t)^\top\right|\right|_F \tag{94}$$

$$= \frac{1}{n}\left|\left|J(\theta_0)\left(J(\theta_0)^\top - J(\theta_t)^\top\right) + \left(J(\theta_0) - J(\theta_t)\right)J(\theta_t)^\top\right|\right|_F \tag{95}$$

$$\leq \frac{1}{n}\left(\left|\left|J(\theta_0)\right|\right|_{op}\left|\left|J(\theta_0)^\top - J(\theta_t)^\top\right|\right|_F + \left|\left|J(\theta_0) - J(\theta_t)\right|\right|_{op}\left|\left|J(\theta_t)^\top\right|\right|_F\right) \tag{96}$$

$$\leq \frac{1}{n}\left(K\sqrt{n}K\left|\left|\theta_0 - \theta_t\right|\right|_2\sqrt{n} + K\left|\left|\theta_0 - \theta_t\right|\right|_2\sqrt{n}K\sqrt{n}\right) \qquad \text{[B.1]} \tag{97}$$

$$= 2K^2\left|\left|\theta_0 - \theta_t\right|\right|_2 \tag{98}$$

$$\leq \frac{6K^3 R_0}{\lambda_{\min}}\frac{1}{\sqrt{n}} \qquad \text{[58]} \tag{99}$$

Bounding the Discrepancy Between the Original and Linearized Networks. Thus far, we have proven both 3 and 4. The only formula we have yet to prove in theorem 2.1 is 2, which bounds the discrepancy between the network outputs in the outputs of its linearized counterpart f^{lin}, restated as follows for convenience:

$$\sup_{t\geq 0}\left|\left|f_t(x) - f_t^{lin}(x)\right|\right|_2 = \mathcal{O}\left(\frac{1}{\sqrt{n}}\right) \tag{100}$$

The paper only proves this for the gradient flow case. Below I derive a bound on this discrepancy for the gradient descent case by utilizing our prior results, and I'm able to get a bound of the desired form $\lesssim {R_0}^2\frac{1}{\sqrt{n}}$. First, recall the basic formulas for f^{lin} and the change in outputs at step $t+1$ compared to step t:

$$f^{lin}(\theta_t) \triangleq f(\theta_t) + J(\theta_0)(\theta_t - \theta_0) \tag{101}$$

$$\implies f^{lin}(\theta_{t+1}) - f^{lin}(\theta_t) = J(\theta_0)(\theta_{t+1} - \theta_t) \tag{102}$$

$$\left|\left|f_{t+1} - f_t\right|\right|_2 \leq \left|\left|J(\theta_t)(\theta_{t+1} - \theta_t)\right|\right|_2 \tag{103}$$

where 103 uses the triangle inequality to discard higher order terms.

$$\left|\left|g^{lin}(t+1) - g(t+1) - (g^{lin}(t) - g(t))\right|\right|_2 = \left|\left|(J(\theta_0) - J(\theta_t))(\theta_{t+1} - \theta_t)\right|\right|_2 \tag{104}$$

$$\leq \left|\left|J(\theta_0) - J(\theta_t)\right|\right|_{op}\left|\left|\theta_{t+1} - \theta_t\right|\right|_2 \tag{105}$$

$$\leq K\sqrt{n}\left|\left|\theta_t - \theta_0\right|\right|_2\left|\left|\theta_{t+1} - \theta_t\right|\right|_2 \qquad \text{[B.1]} \tag{106}$$

$$\leq K\sqrt{n}\left|\left|\theta_t - \theta_0\right|\right|_2\frac{K\eta_0}{\sqrt{n}}\left(1 - \frac{\eta_0\lambda_{\min}}{3}\right)^t R_0 \tag{107}$$

$$= \eta_0 K^2 R_0\left(1 - \frac{\eta_0\lambda_{\min}}{3}\right)^t\left|\left|\theta_t - \theta_0\right|\right|_2 \tag{108}$$

$$\leq \eta_0 K^2 R_0\left(1 - \frac{\eta_0\lambda_{\min}}{3}\right)^t\frac{3KR_0}{\lambda_{\min}}\frac{1}{\sqrt{n}} \tag{109}$$

$$= 3\frac{\eta_0 K^3 {R_0}^2}{\lambda_{\min}}\left(1 - \frac{\eta_0\lambda_{\min}}{3}\right)^t\frac{1}{\sqrt{n}} \tag{110}$$

which gives us an $\mathcal{O}(\frac{1}{\sqrt{n}})$ bound on the *change* in discrepancy between gradient descent steps.

Gradient Flow Case:

1.

$$\frac{d}{dt}\left(\exp\left(\eta_0\hat{\Theta}_0 t\right)\left(g^{lin}(t)-g(t)\right)\right) = \eta_0\hat{\Theta}_0\exp\left(\eta_0\hat{\Theta}_0 t\right)\left(g^{lin}(t)-g(t)\right)+\exp\left(\eta_0\hat{\Theta}_0 t\right)\left(\dot{g}^{lin}(t)-\dot{g}(t)\right) \tag{111}$$

$$= \eta_0\exp\left(\eta_0\hat{\Theta}_0 t\right)\left(\hat{\Theta}_t-\hat{\Theta}_0\right)g(t) \tag{112}$$

$$\int \mathrm{d}t\frac{d}{dt}\left(\exp\left(\eta_0\hat{\Theta}_0 t\right)\left(g^{lin}(t)-g(t)\right)\right) = \eta_0\int_0^t \mathrm{d}s\exp\left(\eta_0\hat{\Theta}_0 s\right)\left(\hat{\Theta}_s-\hat{\Theta}_0\right)g(s) \tag{113}$$

$$\exp\left(\eta_0\hat{\Theta}_0 t\right)\left(g^{lin}(t)-g(t)\right) = \eta_0\int_0^t \mathrm{d}s\exp\left(\eta_0\hat{\Theta}_0 s\right)\left(\hat{\Theta}_s-\hat{\Theta}_0\right)g(s) \tag{114}$$

$$g^{lin}(t)-g(t) = \eta_0\int_0^t \mathrm{d}s\exp\left(\eta_0\hat{\Theta}_0(s-t)\right)\left(\hat{\Theta}_s-\hat{\Theta}_0\right)g(s) \tag{115}$$

2. Let $\lambda_0 > 0$ be the smallest eigenvalue of $\hat{\Theta}_0$. Using the shorthand

$$u(t) \triangleq e^{\lambda_0\eta_0 t}\left|\left|g^{lin}(t)-g(t)\right|\right|_2 \tag{116}$$

$$\alpha(t) \triangleq \eta_0\int_0^t \mathrm{d}s e^{\lambda_0\eta_0 s}\left|\left|\hat{\Theta}_s-\hat{\Theta}_0\right|\right|_{op}\left|\left|g^{lin}(s)\right|\right|_2 \tag{117}$$

$$\beta(t) \triangleq \eta_0\left|\left|\hat{\Theta}_t-\hat{\Theta}_0\right|\right|_{op} \tag{118}$$

we can express the previous step as

$$u(t) \leq \alpha(t)+\int_0^t \mathrm{d}s\beta(s)u(s) \tag{119}$$

$$\leq \alpha(t)\exp\left(\int_0^t \mathrm{d}s\beta(s)\right) \tag{120}$$

where the second step is due to an integral form of the Grönwall's inequality.

3. $\left|\left|g^{lin}(t)\right|\right|_2 \leq e^{-\lambda_0\eta_0 t}\left|\left|g^{lin}(0)\right|\right|_2$
4. Let $\sigma_t = \sup_{0\leq s\leq t}\left|\left|\hat{\Theta}_s-\hat{\Theta}_0\right|\right|_{op}$. Then $\left|\left|g^{lin}(t)-g(t)\right|\right|_2 \lesssim \left(\eta_0 t\sigma_t e^{-\lambda_0\eta_0 t+\sigma_t\eta_0 t}\right)\left|\left|g^{lin}(0)\right|\right|_2$
5. From 59 we know that

$$\sup_t \sigma_t \leq \sup_t\left|\left|\hat{\Theta}_0-\hat{\Theta}_t\right|\right|_F \lesssim R_0\frac{1}{\sqrt{n}} \to 0 \tag{121}$$

as $n_1 = \cdots = n_L = n \to \infty$.

C. Extended Notes and Interpretations for Simon et al. [2021]

Review of the NTK. Consider $\hat{f}_\theta : \mathcal{X} \to \mathbb{R}$. One step of gradient descent on training point (x, y) with small learning rate η

$$\begin{aligned} \ell_\theta(x, y) &\triangleq (\hat{f}_\theta(x) - y)^2 && (122) \\ \theta &\to \theta + \delta\theta && (123) \\ \delta\theta &= -\eta \nabla_\theta \ell_\theta(x, y) && (124) \\ &= -2\eta(\hat{f}_\theta(x) - y)\nabla_\theta \hat{f}_\theta(x) && (125) \end{aligned}$$

Note how the MSE has an elegant interpretation: If the prediction $\hat{f}_\theta(x)$ is larger (smaller) than y, that means we need to change θ s.t. $\hat{f}$ will decrease (increase). Therefore, move in the opposite (same) direction as $\nabla_\theta \hat{f}_\theta(x)$. We are interested in how much that single update changed the predictions of our network on some test input x':

$$\begin{aligned} \hat{f}_{\theta+\delta\theta}(x') &= \underbrace{\hat{f}_\theta(x')}_{\text{original prediction}} + \underbrace{\left\langle \nabla_\theta \hat{f}_\theta(x'), \delta\theta \right\rangle}_{\text{linearized change in pred}} + \mathcal{O}(\delta\theta^2) && (126) \\ &= \hat{f}_\theta(x') - \eta(\hat{f}_\theta(x) - y) \left\langle \nabla_\theta \hat{f}_\theta(x'), \nabla_\theta \hat{f}_\theta(x) \right\rangle + \mathcal{O}(\delta\theta^2) && (127) \\ &= \hat{f}_\theta(x) - \eta(\hat{f}_\theta(x) - y)K(x, x') + \mathcal{O}(\delta\theta^2) && (128) \end{aligned}$$

The parameter update moved θ in the direction of steepest descent in the loss landscape *for the given training input x*. Implications:

- If $\left\langle \nabla_\theta \hat{f}_\theta(x'), \nabla_\theta \hat{f}_\theta(x) \right\rangle = 0$, then the *test prediction is unchanged/unaffected* by the weight update. Consider that $\nabla_\theta \hat{f}_\theta(x)$ is a vector in the same space as θ. If, for example, $\theta \in \mathbb{R}^p$, then there exists a set of $p - 1$ orthonormal vectors v_i that are orthogonal to this direction. In fact, these $p - 1$ vectors span a $p - 1$ dimensional subspace of $\mathbb{R}^p$ for which, if we had simply set $\delta\theta$ to any vector in that subspace, the prediction on the training point x wouldn't have changed at all. There are a large number of changes to the parameters θ we can make that won't alter the predictions of the network on a given input x.
- If $\left\langle \nabla_\theta \hat{f}_\theta(x'), \nabla_\theta \hat{f}_\theta(x) \right\rangle = \left|\left|\nabla_\theta \hat{f}_\theta(x')\right|\right|_2 \left|\left|\nabla_\theta \hat{f}_\theta(x)\right|\right|_2$ (i.e. perfectly parallel), that means the test prediction will be changed by the same exact amount as the training prediction was changed as a result of the weight update. A worst-case scenario of catastrophic forgetting would basically be if the ground truth for x' is $-y$, since that means we just updated our weights in the worst possible direction for improving the test prediction on x'. Note the even bigger implication is that a "perfect" task/distribution for this network is when all inputs $\{x^{(i)}\}$ that share the same target y have identical gradients $\nabla_\theta \hat{f}_\theta(x^{(i)})$, *and* for which any other set of inputs $\{x^{(j)}\}$ that have *different* target values y have orthogonal gradients $\nabla_\theta \hat{f}_\theta(x^{(j)})$ to the others.

Figures of Merit of $\hat{f}$

First, the inner product we'll be using is defined as

$$\langle f, g \rangle \triangleq \frac{1}{M} \sum_{x \in \mathcal{X}} g(x) h(x) \tag{129}$$

$$\textbf{[MSE]} \quad \mathcal{E}^{\mathcal{D}}(f) \triangleq \left\langle f - \hat{f}, f - \hat{f} \right\rangle \quad \text{and} \quad \mathcal{E}(f) \triangleq \mathbb{E}_{\mathcal{D}}\left[\mathcal{E}^{\mathcal{D}}(f)\right] \tag{130}$$

$$\textbf{[Learnability]} \quad \mathcal{L}^{\mathcal{D}}(f) \triangleq \frac{\left\langle f, \hat{f} \right\rangle}{\langle f, f \rangle} \quad \text{and} \quad \mathcal{L}(f) \triangleq \mathbb{E}_{\mathcal{D}}\left[\mathcal{L}^{(D)}(f)\right] \tag{131}$$

The Kernel Eigensystem. NB: authors assume hereafter that $m = 1$ (scalar-output functions). The authors are basically treating the entire input space $\mathcal{X}$ like we usually do for just the training data. Let $M = |\mathcal{X}|$ denote the number of possible inputs x to the network.

Function-Space Perspective.

By definition, any kernel function K is *symmetric* and *positive-semidefinite*: Recall that the definition of a kernel function $K(x, x')$ is that it must be expressible as $\langle \phi(x), \phi(x') \rangle$ for some feature function $\phi : \mathcal{X} \to V$ where $V \subseteq \mathcal{X}$.

1. Clearly, this is symmetric wrt the arguments x, x'.
2. To show it is psd, notice that from the definition we see we can write $K = \Phi \Phi^\top$ for the matrix Φ defined as $\Phi_i \equiv \phi(x^{(i)})$. Therefore, for any function $f : \mathcal{X} \to \mathbb{R}$:

$$\langle f | K | f \rangle = \langle f | \Phi \Phi^\top | f \rangle \tag{132}$$

$$= \left|\left| \Phi^\top | f \rangle \right|\right|_2^2 \geq 0 \tag{133}$$

Recall that any linear Hermitian operator H has a set of orthonormal eigenfunctions that form a basis for the Hilbert space that the operator acts upon[10].

$$\langle K(x, \cdot), \phi_i \rangle = \frac{1}{M} \sum_{x' \in \mathcal{X}} K(x, x') \phi_i(x') = \lambda_i \phi_i(x) \tag{137}$$

which is an equivalent way of saying "K is an operator on functions of $\boldsymbol{x} \in \mathcal{X}$ with eigenfunctions $\{\phi_i\}$

[10] Furthermore, H admits a **spectral decomposition** in this eigenbasis

$$H = \sum_i \lambda_i |\phi_i\rangle \langle \phi_i| \tag{134}$$

$$\implies \forall \psi \quad \langle \psi | H | \psi \rangle = \sum_i \lambda_i \langle \psi \mid \phi_i \rangle \langle \phi_i \mid \psi \rangle = \sum_i \lambda_i \langle \phi_i \mid \psi \rangle^2 \tag{135}$$

which implies that, if $||\psi||_2 = 1$, we have $\langle \psi | H | \psi \rangle \geq \min_i \lambda_i$.

$$\langle \phi_j | K | \phi_i \rangle = \langle \phi_j | \left(\sum_k \lambda_k |\phi_k\rangle \langle \phi_k| \right) |\phi_i\rangle = \lambda_i |\phi_i\rangle \qquad [\langle \phi_k \mid \phi_i \rangle = \delta_{k,i}] \tag{136}$$

s.t. $K\left|\phi_i\right\rangle = \lambda_i \left|\phi_i\right\rangle$. Next, note that we can express both f and $\hat{f}$ in the eigenbasis via

$$\left|f\right\rangle = \sum_{i=1}^{M} v_i \left|\phi_i\right\rangle \tag{138}$$

$$\left|\hat{f}\right\rangle = \sum_{i=1}^{M} \hat{v}_i \left|\phi_i\right\rangle \tag{139}$$

It is straightforward to verify/check that $\left\langle f, \hat{f} \right\rangle = \boldsymbol{v}^\top \hat{\boldsymbol{v}}$. Letting $\boldsymbol{\Phi}(\mathcal{D}) := \phi_i(\boldsymbol{x}^{(j)})$ denote the $M \times n$ matrix of eigenfunctions evaluated at the n training points. Then we can write/define $K(\mathcal{D}, \mathcal{D}) = \boldsymbol{\Phi}^\top(\mathcal{D})\boldsymbol{\Lambda}\boldsymbol{\Phi}(\mathcal{D})$. Plugging this in directly:

$$\hat{f}(x) = K(x, \mathcal{D})K(\mathcal{D}, \mathcal{D})^{-1} f(\mathcal{D}) \tag{140}$$

$$= \begin{bmatrix} \phi_1(x) & \cdots & \phi_M(x) \end{bmatrix} \boldsymbol{\Lambda}\boldsymbol{\Phi} \left(\boldsymbol{\Phi}^\top(\mathcal{D})\boldsymbol{\Lambda}\boldsymbol{\Phi}(\mathcal{D})\right)^{-1} \boldsymbol{\Phi(\mathcal{D})}^\top \boldsymbol{v} \tag{141}$$

$$\langle \phi_i, f \rangle \triangleq \frac{1}{M} \sum_{x \in \mathcal{X}} \phi_i(x) f(x) \tag{142}$$

$$= \frac{1}{M} \sum_{x \in \mathcal{X}} \sum_{i'=1}^{M} \phi_i(x)\phi_{i'}(x) \left(\boldsymbol{\Lambda}\boldsymbol{\Phi} \left(\boldsymbol{\Phi}^\top(\mathcal{D})\boldsymbol{\Lambda}\boldsymbol{\Phi}(\mathcal{D})\right)^{-1} \boldsymbol{\Phi(\mathcal{D})}^\top \boldsymbol{v}\right)_{i'} \tag{143}$$

$$= \frac{1}{M} \sum_{x \in \mathcal{X}} \sum_{i'=1}^{M} \phi_i(x)\phi_{i'}(x)\lambda_{i'} \left(\boldsymbol{\Phi} \left(\boldsymbol{\Phi}^\top(\mathcal{D})\boldsymbol{\Lambda}\boldsymbol{\Phi}(\mathcal{D})\right)^{-1} \boldsymbol{\Phi(\mathcal{D})}^\top \boldsymbol{v}\right)_{i'} \tag{144}$$

$$= \sum_{i'} \lambda_{i'} \left(\boldsymbol{\Phi} \left(\boldsymbol{\Phi}^\top(\mathcal{D})\boldsymbol{\Lambda}\boldsymbol{\Phi}(\mathcal{D})\right)^{-1} \boldsymbol{\Phi(\mathcal{D})}^\top \boldsymbol{v}\right)_{i'} \underbrace{\frac{1}{M} \sum_{x \in \mathcal{X}} \phi_i(x)\phi_{i'}(x)}_{\delta_{i,i'}} \tag{145}$$

$$= \lambda_i \left(\boldsymbol{\Phi} \left(\boldsymbol{\Phi}^\top(\mathcal{D})\boldsymbol{\Lambda}\boldsymbol{\Phi}(\mathcal{D})\right)^{-1} \boldsymbol{\Phi(\mathcal{D})}^\top \boldsymbol{v}\right)_{i} \tag{146}$$

$$= \lambda_i \underbrace{\phi_i(\mathcal{D})^\top}_{1 \times n} \underbrace{\left(\boldsymbol{\Phi}^\top(\mathcal{D})\boldsymbol{\Lambda}\boldsymbol{\Phi}(\mathcal{D})\right)^{-1} \boldsymbol{\Phi(\mathcal{D})}^\top \boldsymbol{v}}_{n \times 1} \tag{147}$$

and therefore

$$\hat{\boldsymbol{v}} = \underbrace{\boldsymbol{\Lambda}\boldsymbol{\Phi}(\mathcal{D}) \left(\boldsymbol{\Phi}^\top(\mathcal{D})\boldsymbol{\Lambda}\boldsymbol{\Phi}(\mathcal{D})\right)^{-1} \boldsymbol{\Phi}^\top(\mathcal{D})}_{\triangleq \boldsymbol{T}^{(\mathcal{D})}} \boldsymbol{v} \tag{148}$$

where $\boldsymbol{T}^{(\mathcal{D})}$ is the **learning transfer matrix**.

Exact Results (2.4).

Lemma 1

(a) $\mathcal{L}^{\mathcal{D})(\phi_i)} = \boldsymbol{T}_{ii}^{\mathcal{D}}$ *and* $\mathcal{L}(\phi_i) = \boldsymbol{T}_{ii}$.
(e) Let $\mathcal{D}_+ = \mathcal{D} \cup x$, *where* $x \in \mathcal{X}$, $x \notin \mathcal{D}$ *is a new data point. Then* $\mathcal{L}^{\mathcal{D}_+}(f) \geq \mathcal{L}^{\mathcal{D}}(f)$

Proof Sketch: Property (e) of Lemma 1

1. Rewrite $\boldsymbol{T}^{\mathcal{D}}$ with $\boldsymbol{\Lambda} \to \boldsymbol{\Lambda}^{1/2}\boldsymbol{\Lambda}^{1/2}$ and observe it is a bunch of products of the $m \times n$ matrix $\boldsymbol{A} \triangleq \boldsymbol{\Lambda}^{1/2}\boldsymbol{\Phi}$:

$$\boldsymbol{T}^{\mathcal{D}} = \boldsymbol{\Lambda}^{\frac{1}{2}}\boldsymbol{A}\left(\boldsymbol{A}^\top\boldsymbol{A}\right)^{-1}\boldsymbol{A}^\top\boldsymbol{\Lambda}^{\frac{1}{2}} = \boldsymbol{\Lambda}^{\frac{1}{2}}(\boldsymbol{A}\boldsymbol{A}^\top)(\boldsymbol{A}\boldsymbol{A}^\top)^{+}\boldsymbol{\Lambda}^{\frac{1}{2}}$$

   Note that the above implies that the $n \times n$ matrix $\boldsymbol{A}^\top\boldsymbol{A}$ is invertible (a consequence of the fact that $n \leq M$ and $\boldsymbol{A}$ has full column rank), but not necessarily $\boldsymbol{A}\boldsymbol{A}^\top$, hence our use of the pseudoinverse.
2. Notice that the effect of appending one more data point is appending one more column to $\boldsymbol{\Phi}$, which we will denote as the M-dimensional vector ξ.
3. The Sherman-Morrison formula tells us how we can evaluate an inverse of the form $(\boldsymbol{B}+uu^\top)^{-1}$ if $\boldsymbol{B}$ is invertible. Here, we'll set $\boldsymbol{B} := \boldsymbol{A}\boldsymbol{A}^\top + \delta\boldsymbol{I}_M$ and $u := \boldsymbol{\Lambda}^{\frac{1}{2}}\xi$. Combining this with the limit definition of the pseudoinverse gives us

$$\boldsymbol{T}^{\mathcal{D}+} = \boldsymbol{T}^{\mathcal{D}} + \lim_{\delta \to 0^+} \delta \frac{\boldsymbol{B}^{-1}\xi\xi^\top\boldsymbol{B}^{-1}}{1 + \xi^\top\boldsymbol{B}^{-1}\xi}$$

4. Since

$$\mathcal{L}^{(\mathcal{D}+)}(f) \propto \boldsymbol{v}^\top\boldsymbol{T}^{(D+)}\boldsymbol{v} \tag{149}$$
$$= \mathcal{L}^{(\mathcal{D})} + \lim_{\delta \to 0^+} \delta\boldsymbol{v}^\top(\cdots)\boldsymbol{v} \tag{150}$$

   and the matrices inside the $(\cdots)$ are psd, we have the desired result.

Deriving a Closed-Form Expression for $\boldsymbol{T}$ (2.5). To motivate the following derivations, recall the resolution of the identity using Dirac notation:

$$\boldsymbol{I}_M = \sum_{i=1}^{M} |\phi_i\rangle\langle\phi_i| = \boldsymbol{\Phi}^\top\boldsymbol{\Phi} \tag{151}$$

where it's important to emphasize that, for now, we should interpret the above solely from the perspective of some abstract Hilbert space with some orthonormal basis of eigenfunctions ϕ_i, and to view $\boldsymbol{\Phi}$ as an associated linear operator. The point is that $\boldsymbol{\Phi}$ is [clearly] a *unitary operator*. Now, if we restrict to the n training points in some dataset $\mathcal{D}$, we should observe that things like the above become an *approximation*, e.g.

$$\langle\phi_i \mid \phi_j\rangle \triangleq \frac{1}{M}\sum_{x \in \mathcal{X}} \phi_i(x)\phi_j(x) = \delta_{i,j} \tag{152}$$
$$\approx \frac{1}{n}\sum_{x \in \mathcal{D}} \phi_i(x)\phi_j(x) \tag{153}$$
$$= \phi_i^\top(\mathcal{D})\phi_j(\mathcal{D}) \tag{154}$$

This motivates re-writing $\boldsymbol{T} \triangleq \mathbb{E}_{\mathcal{D}}\left[\boldsymbol{T}^{(\mathcal{D})}\right]$ as

$$\boldsymbol{T} = \lim_{n \to \infty} \mathbb{E}_{\substack{\boldsymbol{\Phi} \sim \mathbb{R}^{M \times n} \\ \boldsymbol{\Phi}^\top\boldsymbol{\Phi} = \boldsymbol{I}_n}}\left[\boldsymbol{\Lambda}\boldsymbol{\Phi}\left(\boldsymbol{\Phi}^\top\boldsymbol{\Lambda}\boldsymbol{\Phi}\right)^{-1}\boldsymbol{\Phi}^\top\right] \tag{155}$$

and thus we can *approximate* $\boldsymbol{T}$ by computing the above expectation for some reasonably large value of n.

D. Notes for Canatar et al. [2021]

A **reproducing kernel Hilbert space** (RKHS) $\mathcal{H}$ living on $\mathcal{X} \subset \mathbb{R}^D$ is a subset of *square integrable functions*[11] $L_2(\mathcal{X}, p)$ for measure p equipped with an inner product $\langle \cdot, \cdot \rangle_{\mathcal{H}}$ and a kernel K satisfying the **reproducing property**:

$$f(x) = \langle f(\cdot), K(\cdot, x) \rangle_{\mathcal{H}} \qquad (\forall x \in \mathcal{X})(\forall f \in \mathcal{H}) \tag{156}$$

Define the **integral operator** $T_K : L_2(\mathcal{X}, p) \rightarrow L_2(\mathcal{X}, p)$, which is a linear map from functions to functions:

$$T_K[f](x') \triangleq \int p(x) K(x, x') f(x) \mathrm{d}x \tag{157}$$

$$\textbf{[Mercer Decomposition]} \qquad K(x, x') = \sum_{\ell=0}^{\infty} \lambda_\ell \phi_\ell(x) \phi_\ell(x') \tag{158}$$

$$\begin{aligned}
\textbf{[Eigenfunction Property]} \qquad T_K[\phi_\ell](x') &= \int p(x) K(x, x') \phi_\ell(x) \mathrm{d}x && (159)\\
&= \int p(x) \left(\sum_{\ell'=0}^{\infty} \lambda_{\ell'} \phi_{\ell'}(x) \phi_{\ell'}(x') \right) \phi_\ell(x) \mathrm{d}x && (160)\\
&= \sum_{\ell'=0}^{\infty} \lambda_{\ell'} \phi_{\ell'}(x') \int p(x) \phi_{\ell'}(x) \phi_\ell(x) \mathrm{d}x && (161)\\
&= \sum_{\ell'=0}^{\infty} \lambda_{\ell'} \phi_{\ell'}(x') \delta \ell, \ell' && (162)\\
&= \lambda_\ell \phi_\ell(x') && (163)
\end{aligned}$$

A function f is said to be a member of the RKHS $\mathcal{H}$ if and only if $||f||_{\mathcal{H}}^2 < \infty$, where

$$||f||_{\mathcal{H}}^2 = \langle f, f \rangle_{\mathcal{H}} = \sum_{\ell, \ell'} a_\ell a_{\ell'} \langle \phi_\ell, \phi_{\ell'} \rangle_{\mathcal{H}} = \sum_{\ell=0}^{\infty} \frac{1}{\lambda_\ell} a_\ell^2 \tag{164}$$

where the dimension of the RKHS equals the number of nonzero eigenvalues λ_ℓ.

[11] $\int |f(x)|^2 \mathrm{d}x < \infty$