ORF 543 - Deep Learning Theory
Literature Review

Evan Dogariu
November 2, 2022

Background

The field of current deep learning theory research is mainly centered around three important subfields: (1) initialization, (2) feature learning, and (3) training dynamics/phenomenology. (1) deals mainly with understanding the constraints and desired properties of neural networks (NNs) on random initialization, focusing on things like variance selection and initialization schema to make the resulting random NN have useful characteristics. (2) zooms in on the properties of NNs that are conducive to feature learning, focusing on generalizability, embedding functions, the way that NN evaluation and intermediate activations change between data points, etc. (3) focuses on what happens within a NN during optimization, generally through the lens of gradient-based methods. This literature review focuses mostly on (1) and (2), since my own research interests (which are extremely flexible and mobile) currently lie at the intersection of the two (*something along the lines of investigating initialization/parameterization schema for feature learning in* ***autoencoders****, and useful metrics or properties we want them to have during initialization and early training*). An important piece of the literature review that I have not yet completed is to carefully survey current literature on variational and deterministic autoencoders for inspiration about what kinds of properties are desired and attainable at initialization, how to usefully measure feature learning in an autoencoder as opposed to for general architectures, etc. I tried to focus more on the deep learning theory field here, and will figure this out and add it here later. :)

Review of Papers

In this section, I look at recent papers in the interesting and useful subfields to try and get a good grasp of what is going on and how my own research can fit in and complement it. I will start with the more foundational papers and results, and build to more specific and recent results later. I will also try and group by topic (initialization, parameterization, feature learning-focused methods, NTK-focused methods, etc), but a lot of stuff is hard to categorize. I split between NTK and feature learning stuff solely because of the result that in the infinite-width limit, certain parameterizations admit nontrivial kernel methods one or the other, but not both [**1**]. However, lots of NTK methods focus on how to allow finite-width methods to get out of the NTK regime to admit feature learning, and so there is considerable overlap which I think is actually the most interesting, and I discuss that later.

Intro + Infinite Width Limits

Before we can talk about how certain subfields interplay with each other and where my own research can perhaps fit in, it is crucial to look into one of the most important lenses with which to view deep learning theory: the infinite width (or channel, or head, etc.) limit. This limit is natural, as we have empirically observed that overparameterized models with large width perform and generalize very well, indicating that there is something to look at. Also, the math works out to yield some fundamental results that are important in every subfield of deep learning theory. I list a few below.

1. Firstly, Lee et al, 2018 [**2**] show that fully-connected NNs, in the limit as the hidden layer widths go to infinity, converge to **Gaussian Processes** given a prior over the distribution of parameters. The covariance of these Gaussians between data points can be recursively defined, and analytically computed for some selections of nonlinearity. The mathematical simplification that occurs here allows for useful investigation of how different initialization and parameterization paradigms change the way that models behave on initialization and therefore during the training process. We discuss this in the initialization/parameterization section.

2. Jacot et. al, 2018 [**3**] develop a framework for conceptualizing the way that gradient descent occurs in the parameter space via a kernel-based gradient descent in the space of functions. The **Neural Tangent Kernel** (NTK) is a useful construction that describes how update gradient steps look during training, and it has the important property that it is deterministic and fixed throughout training in the infinite width limit. It can be shown via the NTK that training loss is driven down to 0 (we did this in class), and that in the infinite-width limit NN optimization is equivalent to kernel-based linear optimization, which does not admit the feature learning of finite-width models or other parameterization schema. The next result rigorously draws this distinction; however, even though my interests lie on the side of massaging an autoencoder into performing feature learning, the NTK methods and perspective is still invaluable insight into how things evolve during training.

3. Lastly, Yang et. al, 2022 [**1**] rigorously investigate where and how feature learning can happen. In particular, they prove **Dynamical Dichotomy**: a goofily named but foundational result that under any parameterization scheme, infinite-width limits of NNs can either admit feature learning (nontrivial evolution of data-dependent embeddings over training) or be described by kernel gradient descent methods, but not both. So, the NTK parameterization used by Jacot et. al cannot admit feature learning, agreeing with the above results. In doing so, they construct a rigorous and formulaic method to describe arbitrary parameterization schema, as well as how they admit NNs that can either feature learn or adopt an NTK-based description. *Importantly, this result only holds in the infinite-width limit!* Because of this, the result serves a twofold purpose (among many other folds as well): (1) when studying infinite-width characteristics, there is a clear dichotomy that must be worked with, and (2) in finite-width settings, often the interplay between higher-order NTK behaviors and feature learning behaviors is where cool research can be situated.

With these three foundational results/lenses, we can now move to more specific subfields whose explorations are motivated by the above.

Initialization/Parameterization

Before training can occur, a NN must have its parameters initialized via some scheme. The general approach is to sample these scalars from zero-mean Gaussians, and the design of the scheme centers on how we select the variances and how we parameterize the model and learning rate with respect to some scaling parameters, ususally the width. The general idea is that there is some property of the initialized NN that we want to have (in expectation or w.h.p) that constrains our selections in some informative way. For example, the constraint of information propagation (conservation of dot product/covariance, sometimes called signal propagation or geometric information) motivates a certain initialization scheme (Xavier, He, different for different activations). The way that different initialization schema are studied relative to the motivating constraint/goal tends to be through the lens of the infinite-width limit, as this makes the behavior of the models on random initialization describable and able to be worked with. The effects of different initializations and metrics on how feature learning can occur is a rich field as well. One example result in this area comes from Lou et. al, 2022 [**4**], who prove that with the geometric information-conserving initializaiton, large-depth limits of finite-width feed-forward NNs create a nonuniform distribution of geometric information conservation throughout the model, which peaks somewhere in the middle. (Aside: *I think what is most interesting about this paper to me is their systematic treatment of feature learning via parameter freezing, which is only allowing gradient updates to one layer at a time, revealing that with this assumption change of the NN function during early training is controlled by geometric information between the NTK matrix and the data matrix. Simplifications like this simplify the analysis enough to discover cool phenomena that might actually hold in the general case too, and reading papers like this just shows how deep the arsenal of tools and methods can get.*)

A crucial framework with which to think about this subfield is the one of parameterization - how we parameterize the parameters and hyperparameters with respect to certain variables whose growth changes things greatly (namely, the width). For example, the standard parameterization is the landscape in which the standard Xavier, He, etc. initialization schema are derived; this parameterization, however, is not suitable for looking into the NTK, which is when NTK parameterization is used. Another parameterization scheme, Mean Field Parameterization and initialization (Mei et. al, 2018 [**5**]), was developed with the motivation for propagating the maximum amount of mutual information (KL divergence between joint and marginal distributions) of pairs of inputs; it has been shown that this scheme performs well on feature learning in finite-width and infinite-width regimes.

A very important leap in this field of parameterization research came from Yang et. al, 2019 [**1**], which rigorously and thoroughly investigates the consequences of arbitrary parameterization schema. They introduce *abc*-parameterization, which at once describes all of the aforementioned parameterization methods and proves general results (generally in the infinite-limit sense) based on the particular values of the parameters. For example, Yang et. al show that there is a very precise condition on the possible parameters that determines if the resulting scheme will be stable and nontrivial (grow not too fast and not too slow). They also prove a very precise equivalence between parameter choice and whether the infinite-width limit admits feature learning or kernel-based gradient descent

(this result is linked to how parameter choice affects the order of growth of change in intermediate activations). This framework of investigating parameterizations is extremely useful in guiding how an initialization scheme designer balances the desires for feature learning and clean training dynamics. In particular, it draws a line between so-called "active" and "lazy" learning that is precise in the infinite-width regime while still being exceptionally motivating in all regimes for research directions such as my intended one, where we want an autoencoder to be poised to perform quality feature learning.

There is much more to discuss about how the art of initialization/parameterization research goes, as well as the selection of what metric or heuristic or constraint one wants their design to be motivated by. Different schema, such as Standard parameterization, NTK parameterization, Mean Field parameterization, Yang et. al's maximally feature-learning Maximal Update parameterization, etc. are all inspired by some property that we want to explore/exploit/inject into a NN. I think every night about useful motivating characteristics that would be specific to an autoencoder, and I have a couple ideas. The above papers help steer these ideas and also illuminate how to use these constraints to design a methodology; however, there are other places to look for guidance than the parameterization perspective.

NTK/Feature Learning Interplay

Following the work of Jacot et. al, 2018 [**3**] and the introduction of the NTK, it has been extremely useful to understand the way that NNs start and continue to optimize through the lens of kernel-based optimization. Yang et. al's proof of the Dynamical Dichotomy property showed that NTK parameterization and viewpoints in the infinite-width limit do not yield feature learning, but that does not mean that there are not useful things for me to do with NTK-based thinking. In fact, there is a large arm of research devoted to how to bridge the gap between the NTK regime and feature learning in finite-width models. I will enumerate a few below, as these joint ways of thinking are what I find most interesting and would like to immerse myself in more.

We have seen that infinite-width limit NTK kernels do not evolve during training and drive training linearly. Models linearized via the NTK can arbitrarily approximate functions in the RKHS space without any guarantees of how feature approximation works. However, a finite-width viewpoint focused on higher-order NTK dynamics is a lot more exciting: Huang et. al, 2019 [**6**] take the first order gradient descent approximation granted by the NTK methods and expand with higher orders, truncating the infinite expansion to approximate training dynamics to arbitrary precision and developing the Neural Tangent Hierarchy and higher order NTKs. Understanding how the higher order terms, such as the differential of the NTK (dNTK), initialize and develop during training can help one understand the dynamics of how feature learning (in the sense of changing response to the same data) becomes integrated into the training process. A good example of this straddling between NTK methods and feature dynamics exists in the analysis of a method known as QuadNTK [**7**], which uses the second-order NTK along with model randomization in order to allow the quadratic term to dominate, approaching quadratic models and allowing the NN to escape the "kernel regime" and admit feature learning in a nontrivial way. Nichani et. al, 2022 [**8**] take this a step farther, where they apply

substantial regularization (based on the spectrum of the NTK) to a model described by the first two orders, and show that it can split its feature learning into dense and sparse subsets in a nontrivial way. This area of adapting gradient-based expansion methods for understanding training dynamics in regimes where feature learning is possible feels deep and fresh, and that we have barely scratched the surface! I also haven't seen many initialization methods that focus on how the higher order NTK terms start out during training (aside from whether or not they are 0), and so this certainly seems like a region of research that I could find useful and enjoy.

There are also studies into the interplay between these two regimes the other way: investigating how similarly feature-learning-centric paradigms (like Mean Field, finite-width) behave to NTK methods during training. An interesting result comes from Hu et. al, 2020 [**9**], who show that finite-width 2-layer models behave during early training much like linear NTK-based gradient descent, where the training in each layer is boundably close to linear NTK gradient descent at early times.

Autoencoders and Tying it Together

All of the above surveying feels very general and spread out (I read papers from all 3 of the initial 3 field subdivisions, after all). However, throughout my initial paper reading I tried to generally understand these landscapes while honing in on the following question: *how does initialization/parameterization inform feature learning in early training, and what are good metrics for (or qualities of) feature learning that can in turn inform initialization/parameterization schema?* I am also interested in applying this to the particular ML problem of autoencoders, which are NN models attempting to learn an identity operation on a certain unlabeled distribution, with a certain hidden layer of reduced width (equivalently, autoencoders are an enseble of an encoder, which learns a dimensionality reduction mapping, and a decoder that learns its inverse; because they are usually trained jointly, it can also be thought of as a single model with a particular reduced-width hidden layer we want to use for embeddings). I think that autoencoders are a solid reduction of the general problem to a space where there are lots of unique properties/inductive biases of autoencoders that could prove useful. I list a few below:

- The infinite-width limit would look interesting, since it would behave like a normal infinite-width limit if we were to select exactly one of the hidden layers to not include in the limit. Equivalently, we would have an infinite-width encoder whose outputs go through an infinite-width decoder, yielding a very cool and slightly different recursive definition of the Gaussian Process covariance kernel.
- In the spirit of the above point, an infinite-width autoencoder has some similar characteristics of a variational autoencoder (see [**10**]), namely that the latent space is continuous and stochastic, and model output determines some parameterization of a distribution in the latent space subject to some priors. Thought processes like this open up such a research topic to borrowing results and methods from VAEs, which are a pretty well-researched model type.
- Because of the fact that autoencoders should be able to be autoregressive (if the output is the same as the input, you can pass it through the model again to re-encode and re-decode), there is an interesting perspective on a sort of infinite-depth

limit to be had here. However, since the model layers are repeated ad infinitum in this infinite-depth limit, perhaps different feature learning results will spring out of it using the depth limit methods a la [**4**]?

- Because of the encoder/decoder viewpoint of an autoencoder, we desire something a little stronger for feature learning than simply intermediate layer updates of non-trivial magnitude. We can make requests of the form "embedding features Φ_E in the encoder should be correlated to features Φ_D that are learned in the decoder" or something like that in a way that crafts a new constraint/metric for feature learning, which in turn can inform new initialization/parameterization schema.

- This one is a little out there, but perhaps there is some initialization scheme centered about coupled draws for the weights between "corresponding" weights in the encoder and decoder? Of course, we can't simply hard code the decoder to be the inverse of the encoder (or perhaps we wouldn't even want to on initialization?), but there may be things we can do that would set up the training dynamics and feature learning nicely to allow for achievement of any of the motivations/goals mentioned above?

The above intuitions are only really motivated by my own thought and what I already know from theoretical and practical experience with autoencoders and the vibes of the methods we have been using in class and in these papers. I would like to read more into the structured ways that researchers approach autoencoders and learned embedding spaces, perhaps less from the perspective of "feature learning" the way Yang et. al define it and more from the perspective of alignment, uniformity, isotropism, etc. However, for the purposes of this literature review I focused mostly on the papers, which I cite below. Also, I am sorry that I am submitting this so late on the day that the midterm is due. I just really wanted to be as thorough and organized as possible in the writeup, because I plan to use this as a reference and guidance throughout the research process as opposed to simply an assignment to turn in. To be truthful, I don't care about grades as much as feeling like I am doing good research, and so I tried to flesh this all the way out; I apologize if this delay messes anything up. I will keep updating this as I read more and learn more, as well.

Hope you enjoyed! Have a lovely day.

References

[1] Yang, G., & Hu, E.. (2020). Feature Learning in Infinite-Width Neural Networks. https://arxiv.org/pdf/2011.14522.pdf

[2] Lee, J., Bahri, Y., Novak, R., Schoenholz, S., Pennington, J., & Sohl-Dickstein, J.. (2017). Deep Neural Networks as Gaussian Processes. https://arxiv.org/pdf/1711.00165.pdf

[3] Jacot, A., Gabriel, F., & Hongler, C. (2018). Neural Tangent Kernel: Convergence and Generalization in Neural Networks. https://arxiv.org/pdf/1806.07572.pdf

[4] Lou, Y., Mingard, C., Nam, Y., Hayou, S.. (2021). Feature Learning and Signal Propagation in Deep Neural Networks. https://arxiv.org/pdf/2110.11749.pdf

[5] Song Mei, Andrea Montanari, & Phan-Minh Nguyen (2018). A mean field view of the landscape of two-layer neural networks. Proceedings of the National Academy of Sciences, 115(33), E7665-E7671. https://www.pnas.org/doi/10.1073/pnas.1806579115

[6] Jiaoyang Huang, & Horng-Tzer Yau (2019). Dynamics of Deep Neural Networks and Neural Tangent Hierarchy. CoRR, abs/1909.08156. https://arxiv.org/pdf/1909.08156.pdf

[7] Bai, Y., & Lee, J.. (2019). Beyond Linearization: On Quadratic and Higher-Order Approximation of Wide Neural Networks. https://arxiv.org/pdf/1910.01619.pdf

[8] Nichani, E., Bai, Y., & Lee, J.. (2022). Identifying good directions to escape the NTK regime and efficiently learn low-degree plus sparse polynomials. https://arxiv.org/pdf/2206.03688.pdf

[9] Hu, W., Xiao, L., Adlam, B., & Pennington, J.. (2020). The Surprising Simplicity of the Early-Time Learning Dynamics of Neural Networks. https://papers.nips.cc/paper/2020/file/c6dfc6b7c601ac2978357b7a81e2d7ae-Paper.pdf

[10] Kingma, D., & Welling, M.. (2013). Auto-Encoding Variational Bayes. https://arxiv.org/pdf/1312.6114.pdf