Dynamical Mean Field Theory of Kernel Evolution in Wide Neural Networks

Blake Bordelon & Cengiz Pehlevan
John Paulson School of Engineering and Applied Sciences, Center for Brain Science
Harvard University
Cambridge MA, 02138
blake_bordelon@g.harvard.edu, cpehlevan@g.harvard.edu

Abstract

We analyze feature learning in infinite-width neural networks trained with gradient flow through a self-consistent dynamical field theory. We construct a collection of deterministic dynamical order parameters which are inner-product kernels for hidden unit activations and gradients in each layer at pairs of time points, providing a reduced description of network activity through training. These kernel order parameters collectively define the hidden layer activation distribution, the evolution of the neural tangent kernel, and consequently output predictions. We provide a sampling procedure to self-consistently solve for the kernel order parameters.

1 Introduction

Deep learning has emerged as a successful paradigm for solving challenging machine learning and computational problems across a variety of domains [1, 2]. However, theoretical understanding of the training and generalization of modern deep learning methods lags behind current practice. Ideally, a theory of deep learning would be analytically tractable, efficiently computable, capable of predicting network performance and internal features that the network learns, and interpretable through a reduced description involving desirably initialization-independent quantities.

Several recent theoretical advances have fruitfully considered the idealization of *wide neural networks*, where the number of hidden units in each layer is taken to be large. Under certain parameterization, Bayesian neural networks and gradient descent trained networks converge to gaussian processes (NNGPs) [3–5] and neural tangent kernel (NTK) machines [6–8] in their respective infinite-width limits. These limits provide both analytic tractability as well as detailed training and generalization analysis [9–16]. However, in this limit, with these parameterizations, data representations are fixed and do not adapt to data, termed the *lazy regime* of NN training, to contrast it from the *rich regime* where NNs significantly alter their internal features while fitting the data [17, 18]. The fact that the representation of data is fixed renders these kernel-based theories incapable of explaining feature learning, an ingredient which is crucial to the success of deep learning in practice [19, 20]. Thus, alternative theories capable of modeling feature learning dynamics are needed.

Recently developed alternative parameterizations such as the mean field [21] and the μP [22] parameterizations allow feature learning in infinite-width NNs trained with gradient descent. Using the Tensor Programs framework, Yang & Hu identified a stochastic process that describes the evolution of preactivation features in infinite-width μP NNs [22]. In this work, we study an equivalent parameterization to μP with self-consistent dynamical mean field theory (DMFT) and recover the stochastic process description of infinite NNs using this alternative technique. In the same large width scaling, we include a scalar parameter γ_0 that allows smooth interpolation between lazy and rich behavior [17]. We provide a new computational procedure to sample this stochastic process and demonstrate its predictive power for wide NNs.

Machine Learning and the Physical Sciences workshop, NeurIPS 2022.

The present work is a short version of our paper appearing in the main meeting [23]. Our contributions in these works are the following:

1. We develop a path integral formulation of gradient flow dynamics in infinite-width networks in the feature learning regime. Our parameterization includes a scalar parameter γ_0 to allow interpolation between rich and lazy regimes and comparison to perturbative methods.
2. Using a stationary action argument, we identify a set of saddle point equations that the kernels satisfy at infinite-width, relating the stochastic processes that define hidden activation evolution to the kernels and vice versa. We develop a numerical algorithm to solve these equations and show they are predictive of wide networks feature learning dynamics.

Our theory is inspired by self-consistent dynamical mean field theory (DMFT) from statistical physics [24–30]. This framework has been utilized in the theory of random recurrent networks [31–36], tensor PCA [37, 38], phase retrieval [39], and high-dimensional linear classifiers [40–43], but has yet to be developed for deep feature learning. By developing a self-consistent DMFT of deep NNs, we gain insight into how features evolve in the rich regime of network training, while retaining many pleasant analytic properties of the infinite-width limit.

2 Problem Setup and Definitions

Our theory applies to infinite-width networks, both fully-connected and convolutional. For notational ease we will focus on the MLP formalism. For input $\boldsymbol{x}_\mu \in \mathbb{R}^D$, we define the hidden *pre-activation* vectors $\boldsymbol{h}^\ell \in \mathbb{R}^N$ for layers $\ell \in \{1, ..., L\}$ as

$$f_\mu = \frac{1}{\gamma\sqrt{N}} \boldsymbol{w}^L \cdot \phi(\boldsymbol{h}^L_\mu) \,, \quad \boldsymbol{h}^{\ell+1}_\mu = \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell \phi(\boldsymbol{h}^\ell_\mu) \,, \quad \boldsymbol{h}^1_\mu = \frac{1}{\sqrt{D}} \boldsymbol{W}^0 \boldsymbol{x}_\mu, \tag{1}$$

where $\boldsymbol{\theta} = \text{Vec}\{\boldsymbol{W}^0, ..., \boldsymbol{w}^L\}$ are the trainable parameters of the network and ϕ is a twice differentiable activation function. Inspired by previous works on the mechanisms of lazy gradient based training, the parameter γ will control the laziness or richness of the training dynamics [17, 18, 22, 44]. Each of the trainable parameters are initialized as Gaussian random variables with unit variance $W^\ell_{ij} \sim \mathcal{N}(0, 1)$. They evolve under gradient flow $\frac{d}{dt}\boldsymbol{\theta} = -\gamma^2 \nabla_{\boldsymbol{\theta}} \mathcal{L}$. The choice of learning rate γ^2 causes $\frac{d}{dt}\mathcal{L}|_{t=0}$ to be independent of γ. To characterize the evolution of weights, we introduce backpropagation variables $\boldsymbol{g}^\ell_\mu = \gamma\sqrt{N} \frac{\partial f_\mu}{\partial \boldsymbol{h}^\ell_\mu} = \dot{\phi}(\boldsymbol{h}^\ell_\mu) \odot \boldsymbol{z}^\ell_\mu$, where $\boldsymbol{z}^\ell_\mu = \frac{1}{\sqrt{N}} \boldsymbol{W}^{\ell\top} \boldsymbol{g}^{\ell+1}_\mu$ is the *pre-gradient* signal.

The relevant objects to characterize feature learning are feature and gradient kernels for each hidden layer $\ell \in \{1, ..., L\}$, defined as $\Phi^\ell_{\mu\alpha}(t, s) = \frac{1}{N} \phi(\boldsymbol{h}^\ell_\mu(t)) \cdot \phi(\boldsymbol{h}^\ell_\alpha(s)) \,, \quad G^\ell_{\mu\alpha}(t, s) = \frac{1}{N} \boldsymbol{g}^\ell_\mu(t) \cdot \boldsymbol{g}^\ell_\alpha(s)$. From these kernels $\{\Phi^\ell, G^\ell\}_{\ell=1}^L$, we can compute the *Neural Tangent Kernel* $K^{NTK}_{\mu\alpha}(t, s) = \nabla_\theta f_\mu(t) \cdot \nabla_\theta f_\alpha(s) = \sum_{\ell=0}^L G^{\ell+1}_{\mu\alpha}(t, s) \Phi^\ell_{\mu\alpha}(t, s)$, [6] and the dynamics of the network function f_μ

$$\frac{d}{dt} f_\mu(t) = \sum_{\alpha=1}^P K^{NTK}_{\mu\alpha}(t, t) \Delta_\alpha(t) \,, \quad \Delta_\mu(t) = -\frac{\partial}{\partial f_\mu} \mathcal{L}|_{f_\mu(t)}, \tag{2}$$

where we define base cases $G^{L+1}_{\mu\alpha}(t, s) = 1, \Phi^0_{\mu\alpha}(t, s) = K^x_{\mu\alpha} = \frac{1}{D} \boldsymbol{x}_\mu \cdot \boldsymbol{x}_\alpha$. The above expressions demonstrate that knowledge of the temporal trajectory of the NTK on the $t = s$ diagonal gives the temporal trajectory of the network predictions $f_\mu(t)$. Following prior works on infinite-width networks [21, 22, 45, 18], we study the mean field limit $N, \gamma \to \infty \,, \quad \gamma_0 = \frac{\gamma}{\sqrt{N}} = \mathcal{O}_N(1)$. The $\gamma_0 = 0$ limit recovers the static NTK limit [6], while $\gamma_0 > 0$ allows feature learning.

3 Self-consistent DMFT

Next, we derive our self-consistent DMFT in a limit where $t, P = \mathcal{O}_N(1)$. Our goal is to build a description of training dynamics purely based on representations, and independent of weights. Studying feature learning at infinite-width enjoys several analytical properties:

- The kernel order parameters Φ^ℓ, G^ℓ concentrate over random initializations but are dynamical, allowing flexible adaptation of features to the task structure.

- In each layer ℓ, each neuron's preactivation h_i^ℓ and pregradient z_i^ℓ become i.i.d. draws from a distribution characterized by a set of order parameters $\{\Phi^\ell, G^\ell, A^\ell, B^\ell\}$.
- The kernels are defined as self-consistent averages (denoted by $\langle\rangle$) over this distribution of neurons in each layer $\Phi^\ell_{\mu\alpha}(t,s) = \left\langle \phi(h^\ell_\mu(t))\phi(h^\ell_\alpha(s)) \right\rangle$ and $G^\ell_{\mu\alpha}(t,s) = \left\langle g^\ell_\mu(t) g^\ell_\alpha(s) \right\rangle$.

The next section derives these facts from a path-integral formulation of gradient flow dynamics.

3.1 Path Integral Construction

Gradient flow after a random initialization of weights defines a high dimensional stochastic process over initalizations for variables $\{\boldsymbol{h}, \boldsymbol{g}\}$. Therefore, we will utilize DMFT formalism to obtain a reduced description of network activity during training. Generally, we separate the contribution on each forward/backward pass between the initial condition and gradient updates to weight matrix $\boldsymbol{W}^\ell$, defining new stochastic variables $\boldsymbol{\chi}^{\ell+1}_\mu(t) = \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell(0)\phi(\boldsymbol{h}^\ell_\mu(t))$, $\quad \boldsymbol{\xi}^\ell_\mu(t) = \frac{1}{\sqrt{N}} \boldsymbol{W}^\ell(0)^\top \boldsymbol{g}^{\ell+1}_\mu(t)$. We let Z represent the moment generating functional (MGF) for these stochastic fields

$$Z[\{\boldsymbol{j}^\ell, \boldsymbol{v}^\ell\}] = \left\langle \exp\left(\sum_{\ell,\mu} \int_0^\infty dt \left[\boldsymbol{j}^\ell_\mu(t) \cdot \boldsymbol{\chi}^\ell_\mu(t) + \boldsymbol{v}^\ell_\mu(t) \cdot \boldsymbol{\xi}^\ell_\mu(t) \right] \right) \right\rangle_{\{\boldsymbol{W}^0(0), \ldots \boldsymbol{w}^L(0)\}}, \quad (3)$$

Performing integration over possible paths for $\boldsymbol{\chi}^\ell, \boldsymbol{\xi}^\ell$, we show that the MGF Z can be described by set of order-parameters $\{\Phi^\ell, \hat{\Phi}^\ell, G^\ell, \hat{G}^\ell, A^\ell, B^\ell\}$ [23]

$$Z[\{\boldsymbol{j}^\ell, \boldsymbol{v}^\ell\}] \propto \int \prod_{\ell\mu\alpha ts} d\Phi^\ell_{\mu\alpha}(t,s) d\hat{\Phi}^\ell_{\mu\alpha}(t,s) dG^\ell_{\mu\alpha}(t,s) d\hat{G}^\ell_{\mu\alpha}(t,s) dA^\ell_{\mu\alpha}(t,s) dB^\ell_{\mu\alpha}(t,s) \quad (4)$$

$$\times \exp\left(N S[\{\Phi, \hat{\Phi}, G, \hat{G}, A, B, j, v\}] \right),$$

$$S = \sum_{\ell\mu\alpha} \int_0^\infty dt \int_0^\infty ds \left[\Phi^\ell_{\mu\alpha}(t,s)\hat{\Phi}^\ell_{\mu\alpha}(t,s) + G^\ell_{\mu\alpha}(t,s)\hat{G}^\ell_{\mu\alpha}(t,s) - A^\ell_{\mu\alpha}(t,s)B^\ell_{\mu\alpha}(t,s) \right]$$

$$+ \ln \mathcal{Z}[\{\Phi, \hat{\Phi}, G, \hat{G}, A, B, j, v\}], \quad (5)$$

where S is the DMFT action and $\mathcal{Z}$ is a single-site MGF, which defines the distribution of fields $\{\chi^\ell, \xi^\ell\}$ over the neural population in each layer. The order parameters A and B are related to the correlations between feedforward and feedback signals in the network.

3.2 Deriving the DMFT Equations from the Path Integral Saddle Point

As $N \to \infty$, the moment-generating function Z is exponentially dominated by the saddle point of S. The equations that define this saddle point also define our DMFT. We thus identify the kernels that render S locally stationary ($\delta S = 0$). The most important equations are those which define $\{\Phi^\ell, G^\ell\}$

$$\frac{\delta S}{\delta \hat{\Phi}^\ell_{\mu\alpha}(t,s)} = \Phi^\ell_{\mu\alpha}(t,s) + \frac{1}{\mathcal{Z}} \frac{\delta \mathcal{Z}}{\delta \hat{\Phi}^\ell_{\mu\alpha}(t,s)} = \Phi^\ell_{\mu\alpha}(t,s) - \left\langle \phi(h^\ell_\mu(t))\phi(h^\ell_\alpha(s)) \right\rangle = 0,$$

$$\frac{\delta S}{\delta \hat{G}^\ell_{\mu\alpha}(t,s)} = G^\ell_{\mu\alpha}(t,s) + \frac{1}{\mathcal{Z}} \frac{\delta \mathcal{Z}}{\delta \hat{G}^\ell_{\mu\alpha}(t,s)} = G^\ell_{\mu\alpha}(t,s) - \left\langle g^\ell_\mu(t) g^\ell_\alpha(s) \right\rangle = 0, \quad (6)$$

where $\langle\rangle$ denotes an average over the stochastic process induced by $\mathcal{Z}$, which is defined below

$$\{u^\ell_\mu(t)\}_{\mu\in[P], t\in\mathbb{R}_+} \sim \mathcal{GP}(0, \boldsymbol{\Phi}^{\ell-1}) \ , \ \{r^\ell_\mu(t)\}_{\mu\in[P], t\in\mathbb{R}_+} \sim \mathcal{GP}(0, \boldsymbol{G}^{\ell+1}),$$

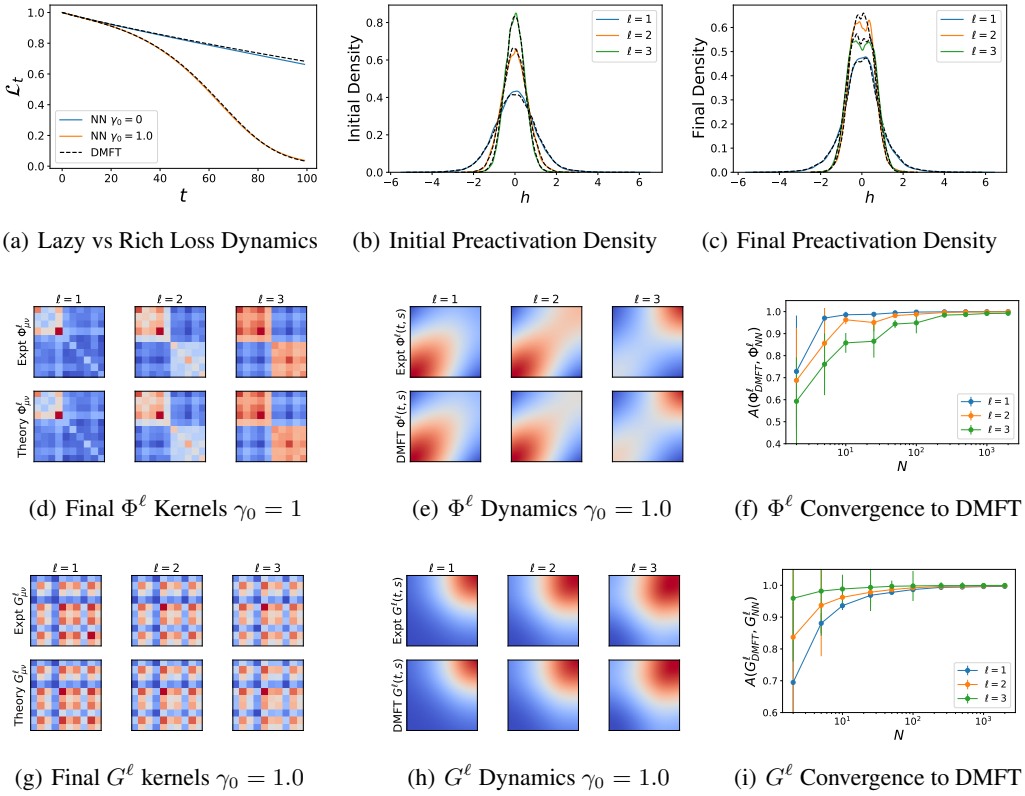
$$h^\ell_\mu(t) = u^\ell_\mu(t) + \gamma_0 \int_0^t ds \sum_{\alpha=1}^P \left[A^{\ell-1}_{\mu\alpha}(t,s) + \Delta_\alpha(s)\Phi^{\ell-1}_{\mu\alpha}(t,s) \right] z^\ell_\alpha(s)\dot{\phi}(h^\ell_\alpha(s)),$$

$$z^\ell_\mu(t) = r^\ell_\mu(t) + \gamma_0 \int_0^t ds \sum_{\alpha=1}^P \left[B^\ell_{\mu\alpha}(t,s) + \Delta_\alpha(s) G^{\ell+1}_{\mu\alpha}(t,s) \right] \phi(h^\ell_\alpha(s)), \quad (7)$$

where we define base cases $\Phi^0_{\mu\alpha}(t,s) = K^x_{\mu\alpha}$ and $G^{L+1}_{\mu\alpha}(t,s) = 1$, $A^0 = B^L = 0$. We see that the fields $\{h^\ell, z^\ell\}$, which represent the single site preactivations and pre-gradients, are implicit

functionals of the mean-zero Gaussian processes $\{u^\ell, r^\ell\}$ which have covariances $\left\langle u^\ell_\mu(t) u^\ell_\alpha(s) \right\rangle = \Phi^{\ell-1}_{\mu\alpha}(t,s)$ and $\left\langle r^\ell_\mu(t) r^\ell_\alpha(s) \right\rangle = G^{\ell+1}_{\mu\alpha}(t,s)$. The other saddle point equations give response functions $A^\ell_{\mu\alpha}(t,s) = \gamma_0^{-1} \left\langle \frac{\delta \phi(h^\ell_\mu(t))}{\delta r^\ell_\alpha(s)} \right\rangle, B^\ell_{\mu\alpha}(t,s) = \gamma_0^{-1} \left\langle \frac{\delta g^{\ell+1}_\mu(t)}{\delta u^{\ell+1}_\alpha(s)} \right\rangle$ which arise due to coupling between the feedforward and feedback signals. We note that, in the lazy limit $\gamma_0 \to 0$, the fields approach Gaussian processes $h^\ell \to u^\ell$, $z^\ell \to r^\ell$.

4 Solving the Self-Consistent DMFT



(a) Lazy vs Rich Loss Dynamics (b) Initial Preactivation Density (c) Final Preactivation Density

(d) Final Φ^ℓ Kernels $\gamma_0 = 1$ (e) Φ^ℓ Dynamics $\gamma_0 = 1.0$ (f) Φ^ℓ Convergence to DMFT

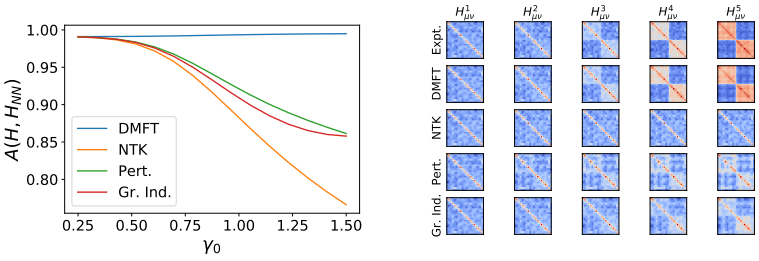
(g) Final G^ℓ kernels $\gamma_0 = 1.0$ (h) G^ℓ Dynamics $\gamma_0 = 1.0$ (i) G^ℓ Convergence to DMFT

Figure 1: Neural network feature learning dynamics is captured by self-consistent dynamical mean field theory (DMFT). (a) Training loss curves on a subsample of $P = 10$ CIFAR-10 training points in a depth 4 ($L = 3$, $N = 2500$) tanh network ($\phi(h) = \tanh(h)$) trained with MSE. Increasing γ_0 accelerates training. (b)-(c) The distribution of preactivations at the beginning and end of training matches predictions of the DMFT. (d) The final Φ^ℓ (at $t = 100$) kernel order parameters match the finite width network. (e) The temporal dynamics of the sample-traced kernels $\sum_\mu \Phi^\ell_{\mu\mu}(t,s)$ matches experiment and reveals rich dynamics across layers. (f) The alignment $A(\boldsymbol{\Phi}^\ell_{DMFT}, \boldsymbol{\Phi}^\ell_{NN})$, defined as cosine similarity, of the kernel $\Phi^\ell_{\mu\alpha}(t,s)$ predicted by theory (DMFT) and width N networks for different N but fixed $\gamma_0 = \gamma/\sqrt{N}$. Errorbars show standard deviation computed over 10 repeats. Around $N \sim 500$ DMFT begins to show near perfect agreement with the NN. (g)-(i) The same plots but for the gradient kernel $\boldsymbol{G}^\ell$. Whereas finite width effects for $\boldsymbol{\Phi}^\ell$ are larger at later layers ℓ since variance accumulates on the forward pass, fluctuations in $\boldsymbol{G}^\ell$ are large in early layers.

The saddle point equations obtained from the field theory discussed in the previous section must be solved self-consistently. By this we mean that, given knowledge of the kernels, we can characterize the distribution of $\{h^\ell, z^\ell\}$, and given the distribution of $\{h^\ell, z^\ell\}$, we can compute the kernels [46, 41]. We use a numerical procedure based on this idea to efficiently solve for the kernels with an alternating Monte-Carlo strategy. The output of the algorithm are the dynamical kernels $\Phi^\ell_{\mu\alpha}(t,s), G^\ell_{\mu\alpha}(t,s), A^\ell_{\mu\alpha}(t,s), B^\ell_{\mu\alpha}(t,s)$, from which any network observable can be computed. We

provide an example of the solution to the saddle point equations compared to training a finite NN in Figure 1. We plot Φ^ℓ, G^ℓ at the end of training and the sample-trace of these kernels through time. We compare kernels of finite width N network to the DMFT predicted kernels using a cosine-similarity alignment metric $A(\mathbf{\Phi}^{DMFT}, \mathbf{\Phi}^{NN}) = \frac{\text{Tr}\, \mathbf{\Phi}^{DMFT} \mathbf{\Phi}^{NN}}{|\mathbf{\Phi}^{DMFT}||\mathbf{\Phi}^{NN}|}$, with agreement at large N.

We can also compare the exact DMFT equations to approximation schemes employed in prior works on wide networks. Specifically, we compare static NTK (NTK), perturbative (Pert.) (leading corrections to feature evolution of size γ_0^2) [47] and gradient independence (Gr. Indep.) (backward pass weights approximated as independent of forward pass) [48]. In Figure 2, we show that only DMFT is accurate over a wide range of feature learning strengths γ_0.



(a) Cosine Similarity with NN (b) Predicted Final Kernels

Figure 2: DMFT is accurate for wide range of γ_0, while other approximations including perturbation theory and gradient independence break down for large γ_0.

5 Broader Impacts

We provided a unifying DMFT derivation of feature dynamics in infinite networks trained with gradient based optimization. Our theory interpolates between lazy infinite-width behavior of a static NTK in $\gamma_0 \to 0$ and rich feature learning. At $\gamma_0 = 1$, our DMFT construction agrees with the stochastic process derived previously with the Tensor Programs framework [22]. Our saddle point equations give self-consistency conditions which relate the stochastic fields to the kernels. These equations are exactly solveable in deep linear networks and can be efficiently solved with a numerical method in the nonlinear case. Comparisons with other approximation schemes show that DMFT can be accurate at a much wider range of γ_0. We believe our framework could be a useful perspective for future theoretical analyses of feature learning and generalization in wide networks, including networks trained with alternative learning rules [49].

Though our DMFT is quite general in regards to the data and architecture, the technique is not entirely rigorous and relies on heuristic physics techniques. Our theory holds in the $T, P = \mathcal{O}_N(1)$ and may break down otherwise; other asymptotic regimes (such as $P/N, T/\log(N) = \mathcal{O}_N(1)$, etc) may exhibit phenomena relevant to deep learning practice [50, 51]. The computational requirements of our method, while smaller than the exponential time complexity for exact solution [22], are still significant for large PT. In Table 1, we compare the time taken for various theories to compute the feature kernels throughout T steps of gradient descent.

Requirements	Width-N NN	Static NTK	Perturbative	Full DMFT
Memory for Kernels	$\mathcal{O}(N^2)$	$\mathcal{O}(P^2)$	$\mathcal{O}(P^4T)$	$\mathcal{O}(P^2T^2)$
Time for Kernels	$\mathcal{O}(PN^2T)$	$\mathcal{O}(P^2)$	$\mathcal{O}(P^4T)$	$\mathcal{O}(P^3T^3)$
Time for Final Outputs	$\mathcal{O}(PN^2T)$	$\mathcal{O}(P^3)$	$\mathcal{O}(P^4)$	$\mathcal{O}(P^3T^3)$

Table 1: Computational requirements to compute kernel dynamics and trained network predictions on P points in a depth N neural network on a grid of T time points trained with P data points for various theories. DMFT is faster and less memory intensive than a width N network only if $N \gg PT$. It is more computationally efficient to compute full DMFT kernels than leading order perturbation theory when $T \ll \sqrt{P}$.

Acknowledgments and Disclosure of Funding

This work was supported by NSF grant DMS-2134157 and an award from the Harvard Data Science Initiative Competitive Research Fund. BB acknowledges additional support from the NSF-Simons Center for Mathematical and Statistical Analysis of Biology at Harvard (award #1764269) and the Harvard Q-Bio Initiative.

References

[1] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep learning*. MIT press, 2016.

[2] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *nature*, 521(7553):436–444, 2015.

[3] Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

[4] Jaehoon Lee, Jascha Sohl-dickstein, Jeffrey Pennington, Roman Novak, Sam Schoenholz, and Yasaman Bahri. Deep neural networks as gaussian processes. In *International Conference on Learning Representations*, 2018.

[5] Alexander G. de G. Matthews, Jiri Hron, Mark Rowland, Richard E. Turner, and Zoubin Ghahramani. Gaussian process behaviour in wide deep neural networks. In *International Conference on Learning Representations*, 2018.

[6] Arthur Jacot, Franck Gabriel, and Clement Hongler. Neural tangent kernel: Convergence and generalization in neural networks. In S. Bengio, H. Wallach, H. Larochelle, K. Grauman, N. Cesa-Bianchi, and R. Garnett, editors, *Advances in Neural Information Processing Systems*, volume 31, pages 8571–8580. Curran Associates, Inc., 2018.

[7] Jaehoon Lee, Lechao Xiao, Samuel Schoenholz, Yasaman Bahri, Roman Novak, Jascha Sohl-Dickstein, and Jeffrey Pennington. Wide neural networks of any depth evolve as linear models under gradient descent. *Advances in neural information processing systems*, 32, 2019.

[8] Sanjeev Arora, Simon S Du, Wei Hu, Zhiyuan Li, Russ R Salakhutdinov, and Ruosong Wang. On exact computation with an infinitely wide neural net. *Advances in Neural Information Processing Systems*, 32, 2019.

[9] Simon Du, Jason Lee, Haochuan Li, Liwei Wang, and Xiyu Zhai. Gradient descent finds global minima of deep neural networks. In *International conference on machine learning*, pages 1675–1685. PMLR, 2019.

[10] B. Bordelon, A. Canatar, and C. Pehlevan. Spectrum dependent learning curves in kernel regression and wide neural networks. *International Conference of Machine Learning*, 2020.

[11] 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):1–12, 2021.

[12] Omry Cohen, Or Malka, and Zohar Ringel. Learning curves for overparametrized deep neural networks: A field theory perspective. *Physical Review Research*, 3(2):023034, 2021.

[13] Arthur Jacot, Berfin Simsek, Francesco Spadaro, Clément Hongler, and Franck Gabriel. Kernel alignment risk estimator: Risk prediction from training data. *Advances in Neural Information Processing Systems*, 33:15568–15578, 2020.

[14] Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborova. Learning curves of generic features maps for realistic datasets with a teacher-student model. In A. Beygelzimer, Y. Dauphin, P. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, 2021.

[15] James B Simon, Madeline Dickens, and Michael R DeWeese. Neural tangent kernel eigenvalues accurately predict generalization. *arXiv preprint arXiv:2110.03922*, 2021.

[16] Zeyuan Allen-Zhu, Yuanzhi Li, and Zhao Song. A convergence theory for deep learning via over-parameterization. In *International Conference on Machine Learning*, pages 242–252. PMLR, 2019.

[17] Lenaic Chizat, Edouard Oyallon, and Francis Bach. On lazy training in differentiable programming. *Advances in Neural Information Processing Systems*, 32, 2019.

[18] Mario Geiger, Stefano Spigler, Arthur Jacot, and Matthieu Wyart. Disentangling feature and lazy training in deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2020(11):113301, 2020.

[19] Tom Brown, Benjamin Mann, Nick Ryder, Melanie Subbiah, Jared D Kaplan, Prafulla Dhariwal, Arvind Neelakantan, Pranav Shyam, Girish Sastry, Amanda Askell, et al. Language models are few-shot learners. *Advances in neural information processing systems*, 33:1877–1901, 2020.

[20] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[21] Song Mei, Andrea Montanari, and Phan-Minh Nguyen. A mean field view of the landscape of two-layer neural networks. *Proceedings of the National Academy of Sciences*, 115(33):E7665–E7671, 2018.

[22] Greg Yang and Edward J Hu. Tensor programs iv: Feature learning in infinite-width neural networks. In *International Conference on Machine Learning*, pages 11727–11737. PMLR, 2021.

[23] Blake Bordelon and Cengiz Pehlevan. Self-consistent dynamical field theory of kernel evolution in wide neural networks. In Alice H. Oh, Alekh Agarwal, Danielle Belgrave, and Kyunghyun Cho, editors, *Advances in Neural Information Processing Systems*, 2022.

[24] Paul Cecil Martin, ED Siggia, and HA Rose. Statistical dynamics of classical systems. *Physical Review A*, 8(1):423, 1973.

[25] C De Dominicis. Dynamics as a substitute for replicas in systems with quenched random impurities. *Physical Review B*, 18(9):4913, 1978.

[26] Haim Sompolinsky and Annette Zippelius. Dynamic theory of the spin-glass phase. *Physical Review Letters*, 47(5):359, 1981.

[27] Haim Sompolinsky and Annette Zippelius. Relaxational dynamics of the edwards-anderson model and the mean-field theory of spin-glasses. *Physical Review B*, 25(11):6860, 1982.

[28] G Ben Arous and Alice Guionnet. Large deviations for langevin spin glass dynamics. *Probability Theory and Related Fields*, 102(4):455–509, 1995.

[29] G Ben Arous and Alice Guionnet. Symmetric langevin spin glass dynamics. *The Annals of Probability*, 25(3):1367–1422, 1997.

[30] Gérard Ben Arous, Amir Dembo, and Alice Guionnet. Cugliandolo-kurchan equations for dynamics of spin-glasses. *Probability theory and related fields*, 136(4):619–660, 2006.

[31] A Crisanti and H Sompolinsky. Path integral approach to random neural networks. *Physical Review E*, 98(6):062120, 2018.

[32] Haim Sompolinsky, Andrea Crisanti, and Hans-Jurgen Sommers. Chaos in random neural networks. *Physical review letters*, 61(3):259, 1988.

[33] Moritz Helias and David Dahmen. *Statistical Field Theory for Neural Networks*. Springer International Publishing, 2020.

[34] Lutz Molgedey, J Schuchhardt, and Heinz G Schuster. Suppressing chaos in neural networks by noise. *Physical review letters*, 69(26):3717, 1992.

[35] M Samuelides and Bruno Cessac. Random recurrent neural networks dynamics. *The European Physical Journal Special Topics*, 142(1):89–122, 2007.

[36] Kanaka Rajan, LF Abbott, and Haim Sompolinsky. Stimulus-dependent suppression of chaos in recurrent neural networks. *Physical review e*, 82(1):011903, 2010.

[37] Stefano Sarao Mannelli, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborova. Passed & spurious: Descent algorithms and local minima in spiked matrix-tensor models. In *international conference on machine learning*, pages 4333–4342. PMLR, 2019.

[38] Stefano Sarao Mannelli, Giulio Biroli, Chiara Cammarota, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Marvels and pitfalls of the langevin algorithm in noisy high-dimensional inference. *Physical Review X*, 10(1):011057, 2020.

[39] Francesca Mignacco, Pierfrancesco Urbani, and Lenka Zdeborová. Stochasticity helps to navigate rough landscapes: comparing gradient-descent-based algorithms in the phase retrieval problem. *Machine Learning: Science and Technology*, 2(3):035029, 2021.

[40] Elisabeth Agoritsas, Giulio Biroli, Pierfrancesco Urbani, and Francesco Zamponi. Out-of-equilibrium dynamical mean-field equations for the perceptron model. *Journal of Physics A: Mathematical and Theoretical*, 51(8):085002, 2018.

[41] Francesca Mignacco, Florent Krzakala, Pierfrancesco Urbani, and Lenka Zdeborová. Dynamical mean-field theory for stochastic gradient descent in gaussian mixture classification. *Advances in Neural Information Processing Systems*, 33:9540–9550, 2020.

[42] Michael Celentano, Chen Cheng, and Andrea Montanari. The high-dimensional asymptotics of first order methods with random data. *arXiv preprint arXiv:2112.07572*, 2021.

[43] Francesca Mignacco and Pierfrancesco Urbani. The effective noise of stochastic gradient descent. *arXiv preprint arXiv:2112.10852*, 2021.

[44] Song Mei, Theodor Misiakiewicz, and Andrea Montanari. Mean-field theory of two-layers neural networks: dimension-free bounds and kernel limit. In *Conference on Learning Theory*, pages 2388–2464. PMLR, 2019.

[45] Lenaic Chizat and Francis Bach. On the global convergence of gradient descent for over-parameterized models using optimal transport. *Advances in neural information processing systems*, 31, 2018.

[46] Alessandro Manacorda, Grégory Schehr, and Francesco Zamponi. Numerical solution of the dynamical mean field theory of infinite-dimensional equilibrium liquids. *The Journal of chemical physics*, 152(16):164506, 2020.

[47] Daniel A Roberts, Sho Yaida, and Boris Hanin. The principles of deep learning theory. *arXiv preprint arXiv:2106.10165*, 2021.

[48] Greg Yang. Tensor programs iii: Neural matrix laws. *arXiv preprint arXiv:2009.10685*, 2020.

[49] Blake Bordelon and Cengiz Pehlevan. The influence of learning rule on representation dynamics in wide neural networks, 2022.

[50] Qianyi Li and Haim Sompolinsky. Statistical mechanics of deep linear neural networks: The backpropagating kernel renormalization. *Physical Review X*, 11(3):031059, 2021.

[51] Stéphane d'Ascoli, Maria Refinetti, and Giulio Biroli. Optimal learning rate schedules in high-dimensional non-convex optimization problems, 2022.

Checklist

1. For all authors...
   (a) Do the main claims made in the abstract and introduction accurately reflect the paper's contributions and scope? [Yes] As described in the abstract and introduction, we provide a dynamical field theory of deep networks based on kernel evolution.
   (b) Did you describe the limitations of your work? [Yes] We have an explicit limitations as the last paragraph of the paper in Section 5.
   (c) Did you discuss any potential negative societal impacts of your work? [N/A] This work is theoretical and is very unlikely to present negative social impacts.
   (d) Have you read the ethics review guidelines and ensured that your paper conforms to them? [Yes]
2. If you are including theoretical results...
   (a) Did you state the full set of assumptions of all theoretical results? [Yes] We describe that our theory holds for NN architectures in the infinite-width $N \to \infty$ limit.
   (b) Did you include complete proofs of all theoretical results? [Yes] All claims made in the main text are supported by derivations in the Appendix.
3. If you ran experiments...
   (a) Did you include the code, data, and instructions needed to reproduce the main experimental results (either in the supplemental material or as a URL)? [Yes] Code to reproduce experimental results is provided in the supplementary material.
   (b) Did you specify all the training details (e.g., data splits, hyperparameters, how they were chosen)? [Yes] We provide details of our experiments in 1
   (c) Did you report error bars (e.g., with respect to the random seed after running experiments multiple times)? [Yes] Mentioned 10 repeats in Figure 1 caption.
   (d) Did you include the total amount of compute and the type of resources used (e.g., type of GPUs, internal cluster, or cloud provider)? [N/A] This is a theory paper
4. If you are using existing assets (e.g., code, data, models) or curating/releasing new assets...
   (a) If your work uses existing assets, did you cite the creators? [N/A]
   (b) Did you mention the license of the assets? [N/A]
   (c) Did you include any new assets either in the supplemental material or as a URL? [N/A]
   (d) Did you discuss whether and how consent was obtained from people whose data you're using/curating? [N/A]
   (e) Did you discuss whether the data you are using/curating contains personally identifiable information or offensive content? [N/A]
5. If you used crowdsourcing or conducted research with human subjects...
   (a) Did you include the full text of instructions given to participants and screenshots, if applicable? [N/A]
   (b) Did you describe any potential participant risks, with links to Institutional Review Board (IRB) approvals, if applicable? [N/A]
   (c) Did you include the estimated hourly wage paid to participants and the total amount spent on participant compensation? [N/A]