Here $\psi(\boldsymbol{x})$ play the role of the infinite-width NTK eigenfunctions, which form a complete basis for square-integrable functions $L^2[p]$. The $\epsilon(\boldsymbol{x})$ function describes a component of y with which is uncorrelated with $\psi(\boldsymbol{x})$. We work in the eigenbasis of features as in (Bordelon et al., 2020), so the covariance given by:

$$\langle \psi_k(\boldsymbol{x})\psi_\ell(\boldsymbol{x})\rangle_{\boldsymbol{x}\sim p(\boldsymbol{x})} = \delta_{k\ell}\lambda_k. \tag{2}$$

The power law structure in the λ_k and $\boldsymbol{w}^*$ entries will lead to power law scalings for the test loss and related quantities.

Student Model. Our student model is motivated by a scenario where a randomly initialized finite-width network is trained in the linearized or lazy regime (Chizat et al., 2019; Jacot et al., 2018). Such training can be described through learning linear combinations of the finite-width NTK features. These features will span a lower-dimensional subspace of the space of square-integrable functions, and relate to infinite-width NTK features in a complicated way.

To model this key aspect, the student model uses a projection of the $\boldsymbol{\psi}(\boldsymbol{x})$ features, $\boldsymbol{A}\boldsymbol{\psi}(\boldsymbol{x})$ where $\boldsymbol{A} \in \mathbb{R}^{N\times M}$. These projected features represent the *finite-width* (i.e. empirical) NTK eigenfunctions. This is motivated by the fact that finite width kernel's features can be linearly expanded in the basis of infinite-width features, because infinite-width kernel eigenfunctions are complete.

Our learned function then has the form:

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{N}}\boldsymbol{w}\cdot \boldsymbol{A}\boldsymbol{\psi}(\boldsymbol{x}). \tag{3}$$

Here, we will interpret N as the model size with the $N\to\infty$ limit recovering original kernel. Similar models were studied in (Maloney et al., 2022; Atanasov et al., 2023).

We will focus on the setting where the elements of $\boldsymbol{A}$ are drawn iid from a distribution of mean zero and variance one. See Appendix B for details on the technical assumptions. The motivations for this choice are (1) tractability and (2) it satisfies the constraint that as $N\to\infty$ the student's kernel approaches the infinite-width kernel $\boldsymbol{\psi}$. In more realistic settings, such as when projecting the eigenfunctions of an infinite-width NTK to a finite-width NTK, the form of the $\boldsymbol{A}$ matrix is generally not known.

Training. The model is trained on a random dataset $\mathcal{D} = \{\boldsymbol{x}_\mu, y_\mu\}_{\mu=1}^P$ of size P with gradient flow on MSE loss

$$\frac{\partial}{\partial t}\boldsymbol{w}(t) = \frac{\sqrt{M}}{P\sqrt{N}}\sum_{\mu=1}^P (y(\boldsymbol{x}_\mu) - f(\boldsymbol{x}_\mu, t))\boldsymbol{A}\boldsymbol{\psi}(\boldsymbol{x}_\mu). \tag{4}$$

We explore extensions (momentum, discrete time, one-pass SGD in Appendix K). We track the test and train loss

$$\begin{aligned}\mathcal{L}(t) &= \mathbb{E}_{\boldsymbol{x}}[(f(\boldsymbol{x},t) - y(\boldsymbol{x}))^2],\\ \hat{\mathcal{L}}(t) &= \frac{1}{P}\sum_{\mu=1}^P (f_\mu(t) - y_\mu)^2.\end{aligned} \tag{5}$$

In small size systems, these losses depend on the precise realization of the data $\mathcal{D}$ and matrix $\boldsymbol{A}$. These two quantities can be viewed as the *disorder*. For large systems, these losses approach a well-defined limit independent of the specific realization of $\mathcal{D}, \boldsymbol{A}$. We will use this fact in the next section when analyzing the model.

3. DMFT for Scaling Laws

We next describe a theoretical approach for characterizing the learning curves for this model. The full details of this approach is detailed in Appendices A, B.

We derive a mean field theory for M, N, P large. We analyze both the (1) proportional regime where $N/M = \nu, P/M = \alpha$ and $M, N, P \to \infty$, and (2) non-proportional regime where $M\to\infty$ first and $N, P \gg 1$. The theories derived in these limits are structurally identical (App. G). [1]

Let $\boldsymbol{\Psi} \in \mathbb{R}^{P\times M}$ with $\boldsymbol{\Psi}_k^\mu = \psi_k(\boldsymbol{x}^\mu)$. Also define $\boldsymbol{\Lambda}_{ij} = \lambda_i\delta_{ij}$.The discrepancy between the target weights and the model's effective weights is

$$\boldsymbol{v}^0 \equiv \boldsymbol{w}^\star - \frac{1}{\sqrt{N}}\boldsymbol{A}^\top \boldsymbol{w}(t). \tag{6}$$

The test loss is then given by $\mathcal{L}(t) = \frac{1}{M}\sum_k \lambda_k v_k^0(t)^2$. The $\boldsymbol{v}^0$ vector has the following dynamics:

$$\frac{d}{dt}\boldsymbol{v}^0(t) = -\left(\frac{1}{N}\boldsymbol{A}^\top\boldsymbol{A}\right)\left(\frac{1}{P}\boldsymbol{\Psi}^\top\boldsymbol{\Psi}\right)\boldsymbol{v}^0(t). \tag{7}$$

Already, we can see that generalization can be limited if $\boldsymbol{A}^\top\boldsymbol{A}$ or $\boldsymbol{\Psi}^\top\boldsymbol{\Psi}$ are low rank as the dynamics will be frozen in the nullspace of $\left(\frac{1}{N}\boldsymbol{A}^\top\boldsymbol{A}\right)\left(\frac{1}{P}\boldsymbol{\Psi}^\top\boldsymbol{\Psi}\right)$. Using DMFT, we characterize this limit by tracking $\boldsymbol{v}^0$ together with the following random vectors:

$$\begin{aligned}\boldsymbol{v}^1(t) &= \frac{1}{\sqrt{M}}\boldsymbol{\Psi}\boldsymbol{v}^0(t), \quad \boldsymbol{v}^2(t) = \frac{1}{\alpha\sqrt{M}}\boldsymbol{\Psi}^\top\boldsymbol{v}^1(t),\\ \boldsymbol{v}^3(t) &= \frac{1}{\sqrt{M}}\boldsymbol{A}\boldsymbol{v}^2(t), \quad \boldsymbol{v}^4(t) = \frac{1}{\nu\sqrt{M}}\boldsymbol{A}^\top\boldsymbol{v}^3(t).\end{aligned} \tag{8}$$

The key summary statistics (also called *order parameters*)

[1] While the proportional limit is exact, the finite size N, P theory will also contain fluctuations across realizations of disorder. When relevant, we show these in experiments by plotting standard deviations over draws of data and projection matrices $\boldsymbol{A}$. This variance decays as $\mathcal{O}(1/P + 1/N)$.

are the correlation functions:

$$C_0(t,s) = \frac{1}{M}\boldsymbol{v}^0(t)^\top \boldsymbol{\Lambda}\boldsymbol{v}^0(s),\ C_1(t,s) = \frac{1}{P}\boldsymbol{v}^1(t)\cdot\boldsymbol{v}^1(s),$$
$$C_2(t,s) = \frac{1}{M}\boldsymbol{v}^2(t)\cdot\boldsymbol{v}^2(s),\ C_3(t,s) = \frac{1}{N}\boldsymbol{v}^3(t)\cdot\boldsymbol{v}^3(s),$$

as well as the response functions:

$$R_1(t,s) = \frac{1}{P}\mathrm{Tr}\left[\frac{\delta \boldsymbol{v}^1(t)}{\delta \boldsymbol{v}^1(s)}\right], R_{2,4}(t,s) = \frac{1}{M}\mathrm{Tr}\left[\frac{\delta \boldsymbol{v}^2(t)}{\delta \boldsymbol{v}^4(s)}\right],$$
$$R_3(t,s) = \frac{1}{N}\mathrm{Tr}\left[\frac{\delta \boldsymbol{v}^3(t)}{\delta \boldsymbol{v}^3(s)}\right], R_{0,2}(t,s) = \frac{1}{M}\mathrm{Tr}\left[\boldsymbol{\Lambda}\frac{\delta \boldsymbol{v}^0(t)}{\delta \boldsymbol{v}^2(s)}\right].$$

Here $\frac{\delta \boldsymbol{v}^i(t)}{\delta \boldsymbol{v}^j(s)}$ is the response of $\boldsymbol{v}^i(t)$ to a kick in the dynamics of $\boldsymbol{v}^j$ at time s. See appendix B.2.1 for details.

The test loss $\mathcal{L}$ and train loss $\hat{\mathcal{L}}$ are related to the time-time diagonal of $C_0(t,s)$ and $C_1(t,s)$ respectively

$$\mathcal{L}(t) = C_0(t,t) + \sigma^2\ ,\ \hat{\mathcal{L}}(t) = C_1(t,t). \tag{9}$$

These collective quantities concentrate over random draws of the disorder (Sompolinsky & Zippelius, 1981). We show that these correlation and response functions satisfy a closed set of integro-differential equations which depend on α, ν which we provide in the Appendices A.2.

Further, we show in Appendix A.3 that the response functions possess a *time-translation invariance* property $R(t,s) = R(t-s)$. This enables exact analysis in the Fourier domain $R(\tau) = \int \frac{d\omega}{2\pi} e^{i\omega\tau}\mathcal{R}(\omega)$. These response functions can then be used to solve for the correlation functions $\{C_0, C_1, C_2, C_3\}$.

To understand the convergence of the learned function f along each eigenfunction of the kernel, we introduce the transfer function[2] for mode k, $H_k(t) \equiv \frac{\partial}{\partial w_k^\star}\left\langle v_k^0(t)\right\rangle$. Our key result is that the Fourier transform of H_k can be simply expressed in terms of the Fourier transforms of R_1, R_3:

$$\mathcal{H}_k(\omega) = \frac{1}{i\omega + \lambda_k \mathcal{R}_1(\omega)\mathcal{R}_3(\omega)}. \tag{10}$$

These functions satisfy the self-consistent equations:

$$\mathcal{R}_1(\omega) = 1 - \frac{1}{P}\sum_k \frac{\lambda_k \mathcal{R}_1(\omega)\mathcal{R}_3(\omega)}{i\omega + \lambda_k \mathcal{R}_1(\omega)\mathcal{R}_3(\omega)},$$
$$\mathcal{R}_3(\omega) = 1 - \frac{1}{N}\sum_k \frac{\lambda_k \mathcal{R}_1(\omega)\mathcal{R}_3(\omega)}{i\omega + \lambda_k \mathcal{R}_1(\omega)\mathcal{R}_3(\omega)}. \tag{11}$$

From these solved response functions $\mathcal{R}_1, \mathcal{R}_3$, we can compute local solutions to the correlation functions' two-variable Fourier transform $\mathcal{C}(\omega,\omega')$ which are independent equations for each pair of ω, ω'. Information about the early dynamics can be extracted from high frequencies $\omega \gg 1$ while information about the late-time limit of the system can be extracted from $\omega, \omega' \to 0$ (App. C, D). For example, for the final test loss,

$$\lim_{t\to\infty} \mathcal{L}(t,\alpha,\nu) = \lim_{\omega,\omega'\to 0} (i\omega)(i\omega')\,\mathcal{C}_0(\omega,\omega'). \tag{12}$$

The full temporal trajectory can be obtained with an inverse Fourier transform of $\mathcal{C}_0(\omega,\omega')$. See Appendix A.4.

4. Results

Our results hold for any λ_k and $w_k^\star$ and we provide some simple analytically solvable examples in Appendix I. However, based on empirical observations of NTK spectral decompositions on realistic datasets (Bordelon & Pehlevan, 2022a; Spigler et al., 2020; Bordelon & Pehlevan, 2022a; Bahri et al., 2021; Maloney et al., 2022), here, we focus on the case of power law features. In this setting, eigenvalues and target coefficients decay as a power law in the index k

$$(w_k^\star)^2\lambda_k \sim k^{-a}\ ,\ \lambda_k \sim k^{-b}. \tag{13}$$

We will refer to a as the *task-power* exponent and b as the *spectral decay* exponent[3]. See Figure 7 (a)-(b) for an example with a Residual CNN on CIFAR-5M.

Test loss power laws. For power law features, the test loss will generally be bottlenecked by either training time t (steps of gradient descent), the size of the training set P, or the size of the model N. We can derive bottleneck scalings from our exact expressions for $\mathcal{L}(t,P,N)$ (Appendix J)[4]:

$$\mathcal{L}(t,P,N) \approx \begin{cases} t^{-(a-1)/b}\ ,\ P,N\to\infty\ , \text{ (Time)} \\ P^{-\min\{a-1,2b\}}\ ,\ t,N\to\infty\ , \text{ (Data)} \\ N^{-\min\{a-1,2b\}}\ ,\ t,P\to\infty\ , \text{ (Model)} \end{cases} \tag{14}$$

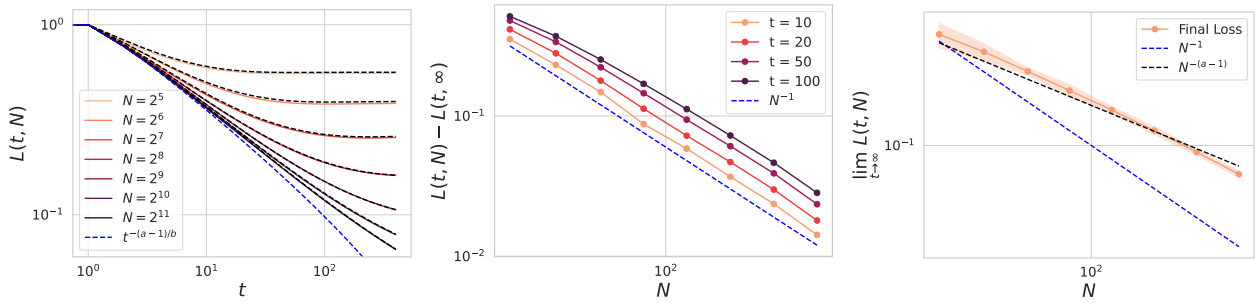
A consequence of this is an *asymmetry in exponent* between the model and data bottlenecks compared to the time bottleneck. We verify this asymmetry in Figure 2.

Bottlenecks as Rank-Constraints All three of the bottleneck scalings arise due to *rank constraints* in the effective dynamics. Heuristically, finite training time or the subsampling of data/features leads to an approximate projection of the target function onto the top $k_\star(t,P,N)$ eigenspace of the infinite-width kernel. The components of the target

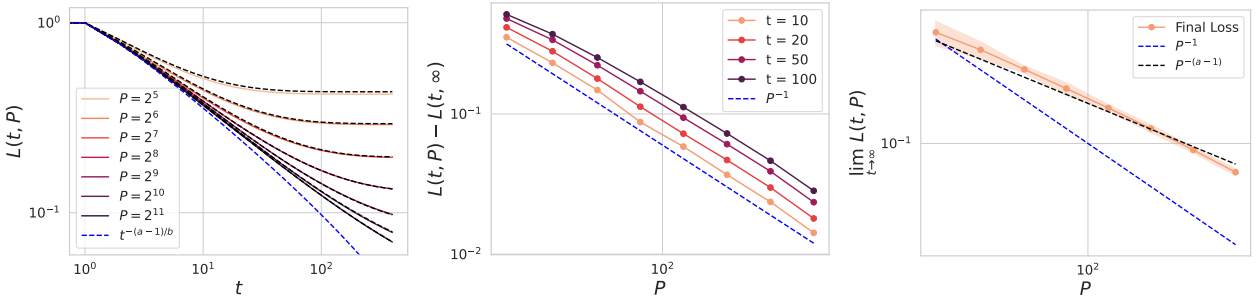
[2] There are dynamical analogues of the mode errors in (Bordelon et al., 2020; Canatar et al., 2021) or learnabilities in (Simon et al., 2021).

[3] These power-law decay rates are also known as source and capacity conditions in the kernel literature (Caponnetto & Vito, 2005; Cui et al., 2021)

[4] The alternative scaling exponents $\mathcal{L} \sim N^{-2b}, P^{-2b}$ occur for very easy tasks which satisfy $a > 2b+1$, but this condition is rarely satisfied in natural data (Appendix J).



(a) $P = 1000$ Test Loss Dynamics (b) Early Time Model Convergence (c) Late Time Model Bottleneck



(d) $N = 1000$ Test Loss Dynamics (e) Early Time Data Convergence (f) Late Time Data Bottleneck

Figure 2. Verification of the various bottleneck scalings for power-law features with $a = 1.5$ and $b = 1.25$. Dashed black lines are DMFT solutions while colors are simulations with standard deviation highlighted. (a) The loss dynamics at large α will be bottlenecked by either time or finite ν. (b) Early in training, the loss converges to its limit as N^{-1} (App. D). (c) At long times, the model's asymptotic loss scales as $N^{-(a-1)}$ (App. C). (d)-(f) The same results but for N and P switched. The model exhibits $1/P$ corrections and early time and power law data bottleneck scalings at late time.

function in the null-space of this projection are not learned. This leads to an approximate test loss of the form

$$\mathcal{L} \approx \sum_{k>k_\star} (w_k^\star)^2 \lambda_k \approx k_\star^{-(a-1)}. \tag{15}$$

For model and data bottlenecks we have that $k_\star \propto N$ and $k_\star \propto P$ respectively (App. J). On the other hand, $k_\star$ for the time bottleneck also depends on the structure of the features through the exponent b. This is because the k-th eigenfeature is learned at a timescale $\tau_k \sim k^b$. At time t, we have learned the first $k_\star \approx t^{1/b}$ modes and the variance in the remaining modes gives $\sim t^{-(a-1)/b}$. In the limit of $t \to \infty$ our data and model bottleneck scalings agree with the resolution and variance-limited scalings studied in (Bahri et al., 2021) as well as prior works on kernels and random feature models (Bordelon et al., 2020; Maloney et al., 2022).

Connection to Online Learning with SGD Many modern deep learning models are trained in an online learning setting where each step of training uses a fresh batch of data to estimate the gradient of the population loss and batches are not reused over multiple steps. Our theoretical methods can also handle this regime. At each step t a fresh minibatch of B examples is used to estimate the gradient. In discrete time with learning rate η this leads to the following DMFT description of $v_k^0(t)$

$$v_k^0(t+1) = v_k^0(t) - \eta u_k^4(t) - \eta \sum_{s\le t} R_3(t,s)[u_k^2(s) + \lambda_k v_k^0(s)] \tag{16}$$

where $u_k^4(t), u_k^2(t)$ are zero-mean Gaussian variables with known covariance (see Appendix K.3). The response function $R_3(t,s)$ satisfies a discrete time analog of Equation (11). The most important observation about this regime is that there is no longer a data bottleneck regime. Rather, the bias component of the test error can only be limited by either training time or model size. The finite batch B introduces SGD noise which introduces an additional variance component to the test loss. We illustrate these results in Figure 3. The $N \to \infty$ limit recovers the results of Bordelon & Pehlevan (2022a) which study online SGD without averaging over a random projection. The continuous time limit of the above expressions obtained from evaluating the theory for small η exactly matches the $P \to \infty$ limit of our gradient flow theory presented in the previous section. We