Following either the MSR or cavity derivation, we obtain an analgous set of limiting DMFT equations defined for integer times $t \in \mathbb{Z}$,

$$\begin{aligned} v_k^0(t+1) &= v_k^0(t) - \eta v_k^4(t) + \delta(t+1) w_k^\star \\ v^1(t) &= u^1(t) + \alpha^{-1} \sum_s R_{0,2}(t,s) v^1(s) \\ v_k^2(t) &= u_k^2(t) + \sum_s R_1(t,s) v_k^0(s) \\ v^3(t) &= u^3(t) + \sum_s R_{2,4}(t,s) v^3(s) \\ v_k^4(t) &= u_k^4(t) + \sum_s R_3(t,s) v_k^2(s) \end{aligned} \tag{133}$$

The delta function in this context is defined as

$$\delta(t+1) = \begin{cases} 1 & t = -1 \\ 0 & \text{else} \end{cases} \tag{134}$$

ensures that the initial condition $v_k^0(0) = w_k^\star$ is satisfied. These iteration equations can be closed for the response functions and correlation functions and solved over $T \times T$ matrices.

Alternatively, we can also solve this problem in an analogous frequency space. Analogous to the Fourier transform method, the equations in discrete time can be closed in terms of the Z-transform

$$v(z) = \sum_{t=-\infty}^{\infty} z^{-t} v(t) \tag{135}$$

Applying this transform gives us the following expression for the v_k^0 fields.

$$v_k^0(z) = \frac{z w_k^\star - \eta u_k^4(z) - \eta \mathcal{R}_3(z) u_k^2(z)}{z - 1 + \eta \lambda_k \mathcal{R}_1(z) \mathcal{R}_3(z)} \equiv \mathcal{H}_k(z) \left[z w_k^\star - \eta u_k^4(z) - \eta \mathcal{R}_3(z) u_k^2(z) \right] \tag{136}$$

Similar to the Fourier case, the final losses can be extracted as the $z \to 1$ limit of these objects.

K.2. Momentum

As mentioned in appendix B, it is straightforward to extend the DMFT treatment beyond just gradient descent dynamics to include a momentum term with momentum β.

We first consider this replacement in continuous time. This requires applying the following replacement:

$$\partial_t v_k^0(t) = -v_k^4(t) \to (\beta \partial_t^2 + \partial_t) v_k^0(t) = -v_k^4(t). \tag{137}$$

This slightly modifies the expressions for the response functions. For example, in Fourier space the response functions become:

$$\begin{aligned} \mathcal{R}_{0,2}(\omega) &= -\frac{1}{M} \sum_k \frac{\lambda_k}{\epsilon + i\omega + \beta(i\omega)^2 + \lambda_k \mathcal{R}_1(\omega) \mathcal{R}_3(\omega)} \mathcal{R}_3(\omega) \\ \mathcal{R}_{2,4}(\omega) &= -\frac{1}{M} \sum_k \frac{\lambda_k}{\epsilon + i\omega + \beta(i\omega)^2 + \lambda_k \mathcal{R}_1(\omega) \mathcal{R}_3(\omega)} \mathcal{R}_1(\omega). \end{aligned} \tag{138}$$

In discrete time, momentum updates can be expressed as

$$\begin{aligned} \boldsymbol{v}^0(t+1) &= \boldsymbol{v}^0(t) - \eta \boldsymbol{b}(t) \\ \boldsymbol{b}(t) &= \boldsymbol{v}^4(t) + \mu \boldsymbol{b}(t-1) \end{aligned} \tag{139}$$

where $\boldsymbol{b}(t)$ is the filtered version of the loss gradient (the $\boldsymbol{v}^4(t)$ field) with momentum coefficient μ and η is the learning rate. The dependence on the $\boldsymbol{b}(t)$ field can be eliminated by turning this into a second order difference equation

$$\boldsymbol{v}^0(t+1) - \boldsymbol{v}^0(t) - \mu\left(\boldsymbol{v}^0(t) - \boldsymbol{v}^0(t-1)\right) = -\eta \boldsymbol{v}^4(t). \tag{140}$$

Again, the final result can be expressed in terms of the Z-transformed transfer functions $\mathcal{H}_k(z)$ which have the form

$$\mathcal{H}_k(z) = \frac{1}{z - 1 - \mu + \mu z^{-1} + \eta \mathcal{R}_1(z)\mathcal{R}_3(z)}. \tag{141}$$

K.3. One Pass SGD

In this section we derive online SGD with projected features. At each step a random batch of B samples are collected (independent of previous samples), giving a matrix $\boldsymbol{\Psi}(t) \in \mathbb{R}^{B\times M}$ of sampled features. The update at step t is

$$\boldsymbol{v}^0(t+1) = \boldsymbol{v}^0(t) + \eta \left(\frac{1}{N}\boldsymbol{A}^\top \boldsymbol{A}\right)\left(\frac{1}{B}\boldsymbol{\Psi}(t)^\top \boldsymbol{\Psi}(t)\right)\boldsymbol{v}^0(t). \tag{142}$$

The DMFT limit gives the following statistical description of the fields, which decouple over time for the $v^1(t), v_k^2(t)$ but remain coupled across time for $v^3(t), v_k^4(t)$

$$\begin{aligned}
v^1(t) &= u^1(t)\ ,\ u^1(t) \sim \mathcal{N}(0, C_0(t,t)\delta(t-s)),\\
v_k^2(t) &= u_k^2(t) + \lambda_k v_k^0(t)\ ,\ u_k^2(t) \sim \mathcal{N}\left(0, \frac{1}{B}\lambda_k C_1(t,t)\delta(t-s)\right)\\
v^3(t) &= u^3(t) + \frac{1}{N}\sum_s R_{2,4}(t,s)v^3(s)\ ,\ u^3(t) \sim \mathcal{N}(0, C_2(t,s))\\
v_k^4(t) &= u_k^4(t) + \sum_s R_3(t,s)v_k^2(s)\ ,\ u_k^4(t) \sim \mathcal{N}\left(0, \frac{1}{N}C_3(t,s)\right)\\
v_k^0(t+1) &= v_k^0(t) - \eta v_k^4(t).
\end{aligned} \tag{143}$$

This system cannot exhibit overfitting effects as we have the statistical equivalence between the covariance of v^1 and the test loss:

$$\hat{\mathcal{L}}(t) = \left\langle v^1(t)^2 \right\rangle = \left\langle u^1(t)^2 \right\rangle = C_0(t,t) = \mathcal{L}(t) \tag{144}$$

We note that this is very different than the case where data is reused at every step, which led to a growing gap between train and test loss as we derive in Appendix E.

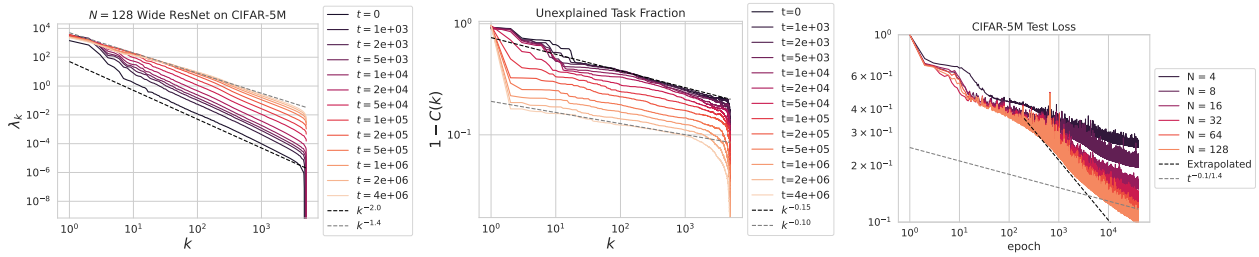
We visualize some example results for one-pass SGD with power law features in Figure 3. While we see that the same scaling laws with t and N hold, the dependence on batchsize B is much weaker: the model never reaches an asymptote that scales with B but rather experiences SGD noise that scales with η/B for learning rate η.

We summarize the key similarities and differences between the one-pass SGD and multi-pass batch GD settings

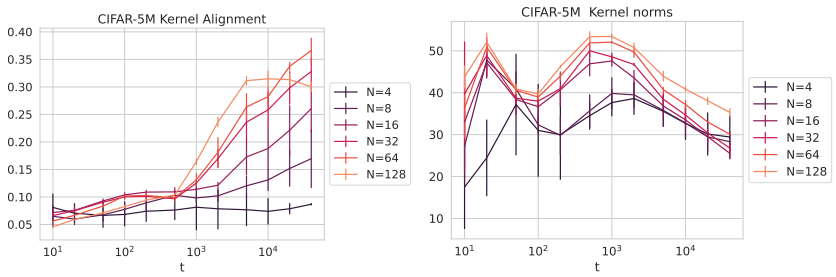
1. If the learning rate is small and a continuous time limit of the dynamics is taken, then the SGD dynamics will agree with the $P \to \infty$ limit of our full batch gradient flow theory. This is a setting where finite data and SGD noise are negligible.
2. If learning rate is non-negligible and batch size is finite, then SGD noise cannot be neglected and the SGD dynamics will be different than full pass GD. The SGD dynamics will be described by a discrete time DMFT given above.
3. In general, the multi-pass version of the theory can have a train loss and test loss gap while the SGD theory never has a gap between training and test loss.
4. The SGD test loss can be limited by t, N, but the effect of finite batch size is basically some additive variance in the model outputs. Finite dataset size in the full batch GD can lead to a bottleneck scaling law (like $L \sim P^{-(a-1)}$).

L. Kernel Analysis of Feature Learning Networks

In Section 5.1, we observed that feature learning networks can achieve better loss and compute-optimal scaling. In such settings, it may be useful to observe the *after kernel*, namely the NTK at the end of training. This object can often shed



(a) Final NTK Spectra for Varying Widths (b) Final NTK Task-Power Decay (c) Losses for Feature Learning Networks



(d) (e)

Figure 12. a) The observed power law spectrum on a held out test set of the after-kernel for a width $N = 128$ ResNet trained on CIFAR-5m. Early on in training, the spectrum flattens quite rapidly. At later times, the spectral decay remains relatively constant. b) The fraction of the task unexplained, as defined in Equation 20. Throughout training, the top eigenmode of the after-kernel explains more and more of the task. c) The test loss of the network. We see that the observed scaling of this quantity is faster than that predicted from analyzing the after-kernel. d) The kernel-target alignment of the after kernel improves throughout training time. The error bars here denote different ensemble members. Their relatively small size implies that the kernel trajectory is relatively deterministic over different initialization seeds. e) The norm of the after-kernel throughout training is relatively constant for this task.

insight into the structure of the learned network function (Atanasov et al., 2022; Long, 2021) and its generalization. In some cases, it has been observed that the final kernel stabilizes during the course of training (Fort et al., 2020), potentially allowing one to potentially deduce scaling laws from the spectrum and task-model alignment of this after-kernel, though other papers have observed contrary results (Vyas et al., 2022).

Motivated by this, we study the NTKs of the finite-width networks trained for 64 epochs with the animate-inanimate CIFAR-5m discrimination task. We observe in Figure 12 a) that the spectrum becomes flatter, with a decay exponent of close to 1.4 down from 2.0 for the initial kernel.

The fraction of the task power unexplained is also observed to have a lower exponent in Figure 12 b), however there is also the presence of a low rank spike indicative of the kernel aligning to this discrimination tasks.

From these scalings we can obtain the a and b exponents and get a prediction for the scaling of the test loss. We plot this in grey in Figure 12 c). The observed scaling (in black) is much better than that predicted by the after-kernel. This is an indication the the after kernel continues evolving in this task, improving the scaling exponent of the test loss.

The kernel-target alignment (Cortes et al., 2012), as measured by

$$A = \frac{\boldsymbol{y}^\top \boldsymbol{K} \boldsymbol{y}}{\boldsymbol{y}^\top \boldsymbol{y} |\boldsymbol{K}|_F}, \tag{145}$$

is plotted in 12 d). Here $\boldsymbol{y}$ is the target labels on a held-out test set, and $\boldsymbol{K}$ is the gram matrix of the after-kernel on this test set. We indeed observe a consistent increase in this quantity across time. This gives an indication that understanding the evolution of the after-kernel will be useful