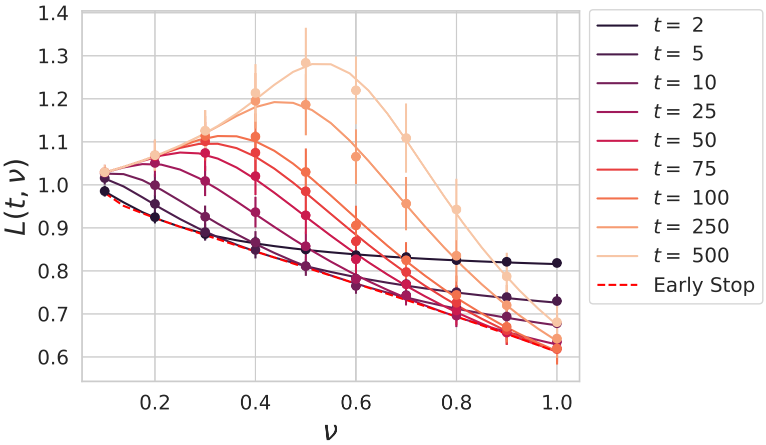

Figure 10. The white bandlimited model ($\lambda_k = 1$) with $\alpha = 0.8$ and varying model size ν with no explicit noise $\sigma = 0$ exhibits double descent at late time. Optimal early stopping, like optimal regularization, recovers monotonic scaling with ν.

Writing $\mathcal{R}_1 = 1 - \frac{\nu}{\alpha}(\mathcal{R}_3 - 1)$ allows us to solve for $\mathcal{R}_3$ exactly:

$$\mathcal{R}_3 \left(i\omega + \mathcal{R}_3 (1 + \frac{\nu}{\alpha}(\mathcal{R}_3 - 1)) \right) = i\omega + \mathcal{R}_3(1 + \frac{\nu}{\alpha}(\mathcal{R}_3 - 1)) - \frac{1}{\nu} \left(\mathcal{R}_3 (1 + \frac{\nu}{\alpha}(\mathcal{R}_3 - 1)) \right) . \tag{118}$$

This is a cubic equation that can be solved for $\mathcal{R}_3$ as a function of ω. In the limit of $\alpha \to \infty$ this simplifies to:

$$\begin{aligned} &\mathcal{R}_3 i\omega + \mathcal{R}_3^2 = i\omega + \left(1 - \frac{1}{\nu}\right) \mathcal{R}_3 \\ &\Rightarrow \mathcal{R}_3 = \frac{1}{2}[(1 - \nu^{-1} - i\omega) + \sqrt{(1 - \nu^{-1} - i\omega)^2 + 4i\omega}]. \end{aligned} \tag{119}$$

I.2. Timescale Corrections in The Small ν Regime

By expanding the above in the limit of small ν we get that $\mathcal{R}_3$ goes as

$$\mathcal{R}_3 \sim \frac{i\omega}{\nu^{-1} - 1 + i\omega} \ , \ \nu \to 0 \tag{120}$$

From this approximate response function, we find that the transfer function takes the form

$$\begin{aligned} H(\tau) &= \int \frac{d\omega}{2\pi} \frac{e^{i\omega\tau}}{i\omega + \frac{i\omega}{\nu^{-1} - 1 + i\omega}} = \int \frac{d\omega}{2\pi} \frac{(\nu^{-1} - 1 + i\omega) e^{i\omega\tau}}{i\omega \left[\nu^{-1} + i\omega\right]} \\ &= (1 - \nu) + \nu e^{-\tau/\nu}, \end{aligned} \tag{121}$$

where in the last line, we used the residue theorem. We note that in this perturbative approximation that this transfer function is always greater than the transfer function at $\nu \to \infty$ which is $e^{-\tau}$. Thus finite ν leads to higher bias in this regime. We define bias and variance precisely in Appendix H.2.

I.3. Timescale corrections in fully expressive regime $\nu > 1$

For $\nu \gg 1$, we can approximate $R_3(\omega) \sim 1 - \nu^{-1}(1 + i\omega)^{-1}$, we have

$$\begin{aligned} H(\tau) &\sim \int \frac{d\omega}{2\pi} \frac{e^{i\omega\tau}}{i\omega + 1 - \nu^{-1}(i\omega + 1)^{-1}} = \int \frac{d\omega}{2\pi} \frac{e^{i\omega\tau}(1 + i\omega)}{(i\omega + 1 - \nu^{-1/2})(i\omega + 1 + \nu^{-1/2})} \\ &= \frac{1}{2} e^{-\tau(1 + \sqrt{\nu})} + \frac{1}{2} e^{-\tau(1 - \sqrt{\nu})} = e^{-\tau} \cosh\left(\tau/\sqrt{\nu}\right) \end{aligned} \tag{122}$$

where we used the residue theorem after closing the contour in the upper half-plane. In Figure 11, we show that this perturbative approximation does capture a slowdown in the dynamics for large but finite ν.

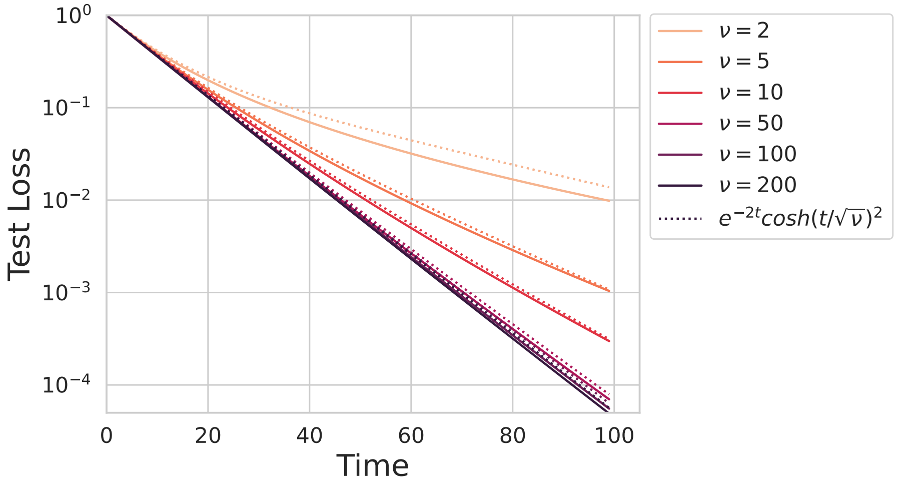

Figure 11. Slower timescales in the $\nu > 1$ regime for white bandlimited features.

J. Power-Law Bottleneck Scalings

In this section we calculate the scaling of the loss with the various limiting resources (time, model size, and data) when using power law features. Since the power-law features give a trace class kernel (*i.e.* $\sum_{k=1}^{\infty} \lambda_k < \infty$), we use the non-proportional limit formalism in Appendix G, which gives an expression for $\mathcal{L}(t, N, P)$ with M already considered infinite. While the resulting expressions are not a formal proportional thermodynamic limit and finite N, P corrections exist in the form of fluctuations from one random realization of the system to another. These corrections decay rapidly enough at finite N, P for this mean field theory to be accurate and descriptive in realistic systems (Bordelon et al., 2020; Simon et al., 2023; Cheng & Montanari, 2022). We plot this variability of random finite size experiments as highlighted standard deviations in the main text figures.

J.1. Time Bottleneck

The time bottleneck is defined as the limiting dynamics in the absence of any model or data finite size effects. To eliminate those effects, we simply study the $\alpha, \nu \to \infty$ limit

$$\mathcal{L}_\infty(t) = \lim_{P,N\to\infty} \mathcal{L}(t, P, N). \tag{123}$$

In this limit, the response functions simplify to $\mathcal{R}_1(\omega)\mathcal{R}_3(\omega) \to 1$ so that

$$\mathcal{H}_k(\omega) = \frac{1}{i\omega + \lambda_k} \implies H_k(\tau) = e^{-\lambda_k \tau}\Theta(\tau). \tag{124}$$

Further, in this limit, we have that $C_0(t, s) = \frac{1}{M}\sum_k \lambda_k H_k(t) H_k(s) (w_k^\star)^2$ since all the variance terms (which depend on ν^{-1}, α^{-1}) drop out. Thus we have the following loss at time t,

$$\mathcal{L}(t) = \sum_k \lambda_k (w_k^\star)^2 e^{-2\lambda_k t} \sim \int_1^\infty dk k^{-a} \exp\left(-2k^{-b} t\right) \sim t^{-(a-1)/b}. \tag{125}$$

where the final scaling with time can be obtained through either change of variables or steepest descent methods (Bordelon & Pehlevan, 2022a).

J.2. Model Bottleneck

In this section we take $\alpha, t \to \infty$. This leaves us with the following equation for $r \equiv \lim_{\omega \to 0} (i\omega)^{-1} \mathcal{R}_3(\omega)$.

$$N = \sum_k \frac{\lambda_k r}{\lambda_k r + 1} \approx \int_1^\infty \frac{dk}{k^b / r + 1} \approx r^{1/b} \implies r \approx N^b. \tag{126}$$

Now, the large time limit of the transfer functions $H_k(\tau)$ can be obtained from the final-value theorem

$$\lim_{t \to \infty} H_k(\tau) = \lim_{\omega \to 0} \frac{i\omega}{i\omega + \lambda_k r i\omega} = \frac{1}{1 + \lambda_k r}. \tag{127}$$

Now, integrating over the eigenvalue density to get the total loss (and disregarding prefactors)

$$\begin{aligned} \mathcal{L}(t) &\sim \int_1^\infty dk \, \frac{k^{-a}}{(1 + k^{-b} r)^2} \\ &\approx \frac{1}{r^2} \int_1^N dk \, k^{2b-a} + \int_N^\infty k^{-a} \\ &= \frac{1}{2b + 1 - a} [N^{-(a-1)} - N^{-2b}] + \frac{1}{a-1} N^{-(a-1)} \\ &\sim N^{-\min\{a-1, 2b\}} \end{aligned} \tag{128}$$

For difficult tasks where $a - 1 < 2b$, we thus expect a powerlaw scaling of the form $\mathcal{L} \sim N^{-(a-1)}$ in this regime.

J.3. Data Bottleneck

In this section we take $\nu, t \to \infty$. This leaves us with the following equation for $r \equiv \lim_{\omega \to 0} (i\omega)^{-1} \mathcal{R}_1(\omega)$.

$$P = \sum_k \frac{\lambda_k r}{\lambda_k r + 1} \approx \int_1^\infty \frac{dk}{k^b / r + 1} \approx r^{1/b} \implies r \approx P^b. \tag{129}$$

Now, the large time limit of the transfer functions $H_k(\tau)$ can again be obtained from the final-value theorem

$$\lim_{t \to \infty} H_k(\tau) = \lim_{\omega \to 0} \frac{i\omega}{i\omega + \lambda_k r i\omega} = \frac{1}{1 + \lambda_k r} \tag{130}$$

Now, integrating over the eigenvalue density to get the total loss gives

$$\begin{aligned} \mathcal{L}(t) &\sim \int_1^\infty dk \, \frac{k^{-a}}{(1 + k^{-b} r)^2} \\ &\approx \frac{1}{r^2} \int_1^P dk \, k^{2b-a} + \int_P^\infty k^{-a} \\ &= \frac{1}{2b + 1 - a} [P^{-(a-1)} - P^{-2b}] + \frac{1}{a-1} P^{-(a-1)} \\ &\sim P^{-\min\{a-1, 2b\}} \end{aligned} \tag{131}$$

For difficult tasks with $a < 2b + 1$, the loss will therefore scale as $P^{-(a-1)}$ in this data-bottleneck regime.

K. Optimization Extensions

K.1. Discrete Time

In this section, we point out that DMFT can also completely describe discrete time training as well. In this section we consider discrete time gradient descent with learning rate η

$$\begin{aligned} &\boldsymbol{v}^0(t+1) = \boldsymbol{v}^0(t) - \eta \boldsymbol{v}^4(t) \\ &\boldsymbol{v}^4(t) = \frac{1}{\nu \sqrt{M}} \boldsymbol{A}^\top \boldsymbol{v}^3(t) \, , \, \boldsymbol{v}^3(t) = \frac{1}{\sqrt{M}} \boldsymbol{A} \boldsymbol{v}^2(t) \\ &\boldsymbol{v}^2(t) = \frac{1}{\alpha \sqrt{M}} \boldsymbol{\Psi}^\top \boldsymbol{v}^1(t) \, , \, \boldsymbol{v}^1(t) = \frac{1}{\sqrt{M}} \boldsymbol{\Psi} \boldsymbol{v}^0(t) \end{aligned} \tag{132}$$