E.2. Low Frequency Range/Late Time

At late time/low frequency, as we showed in Appendix C, the behavior of the C_1 correlation function depends on whether the model is over-parameterized or under-parameterized. In the overparameterized case, the asymptotic train loss is zero while the asymptotic test loss is nonzero. In the underparameterized case, we have a limiting value for both the test and train loss which can be computed from the expressions in Appendix C.

F. Timescale/Eigenvalue Density Interpretation

We can use an alternative interpretation of the Fourier transforms derived in previous sections to obtain the timescale density for the dynamics. Since this is a linear model defined by an effective matrix $\frac{d}{dt}\boldsymbol{v}^0 = -\left(\frac{1}{N}\boldsymbol{A}^\top \boldsymbol{A}\right)\left(\frac{1}{P}\boldsymbol{\Psi}^\top \boldsymbol{\Psi}\right)\boldsymbol{v}^0$, this is equivalent to computing the eigenvalue density. We start by expanding the transfer function for mode k in the basis of exponentials

$$H_k(t) = \int_0^\infty du\, \rho_k(u)\, e^{-ut}. \tag{88}$$

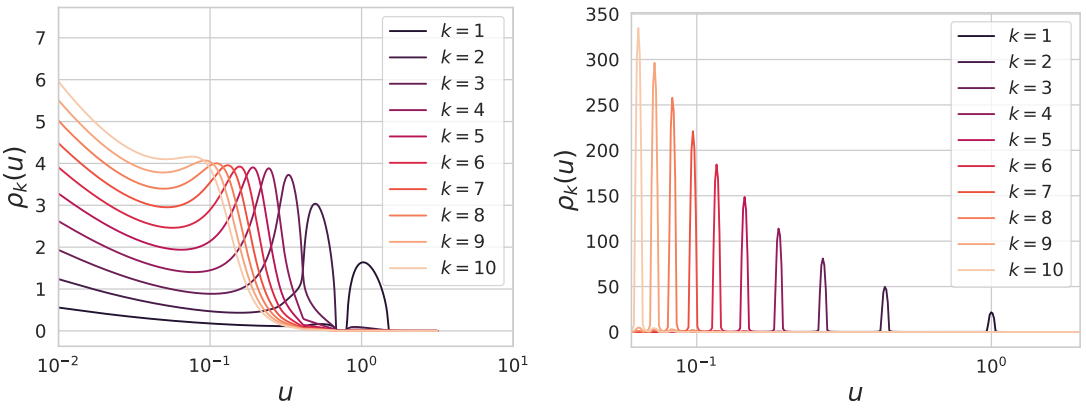
We allow for Dirac-delta masses at $u = 0$ which correspond to the constant (unlearnable) components. Next, we note that the Fourier transform has the form

$$\mathcal{H}_k(\omega) = \int_{-\infty}^\infty dt\, e^{-i\omega t} H_k(t) = \int_0^\infty du\, \rho_k(u) \int_{-\infty}^\infty dt\, e^{-(u+i\omega)t} = \int_0^\infty du\, \frac{\rho_k(u)}{i\omega + u}. \tag{89}$$

We can recover the density $\rho_k(s)$ by using the Sokhotski–Plemelj theorem $\frac{1}{\pi}\text{Im}\frac{1}{-i\epsilon+u-s} = \delta(u-s)$ which gives us

$$\rho_k(u) = \lim_{\epsilon\to 0} \frac{1}{\pi}\text{Im}\, \mathcal{H}_k(iu - \epsilon). \tag{90}$$

This allows us to interpret the spread of timescales from the random sampling of data and the random projection $\boldsymbol{A}$. In the limit of $\alpha, \nu \to \infty$ we have $\rho_k(s) = \delta(s - \lambda_k)$ but for finite α, ν the density spreads out. We visualize these densities for power law features in Figure 8.



(a) $N = 50, P = 100$ (b) $N = 10000, P = 10000$

Figure 8. Timescale (eigenvalue) densities for each transfer function $H_k(\tau)$ with power law features with $b = 1.2$. For limited N, P there is a significant spread of timescales for each mode. For $N, P \to \infty$ the density converges to a Dirac mass at $u = \lambda_k$.

F.1. Recovering the Marchenko-Pastur Law from DMFT Response Functions

To further illustrate the validity of this perspective, we show that it is possible to recover known random matrix theory results using this technique. To illustrate this, we study the case where $\lambda_k = 1$ and take $\nu \to \infty$. In this case, we have the coupled equations

$$\mathcal{H}(\omega) = \frac{1}{i\omega + \mathcal{R}_1(\omega)}\ ,\ \mathcal{R}_1(\omega) = 1 - \frac{1}{\alpha}\mathcal{R}_1(\omega)\mathcal{H}(\omega) \tag{91}$$

Combining these equations gives the single equation

$$\mathcal{H}(\omega) = \frac{1}{i\omega + \frac{\alpha}{\alpha + \mathcal{H}(\omega)}} \ , \implies i\omega\mathcal{H}(\omega)^2 + (\alpha i\omega + \alpha - 1)\mathcal{H}(\omega) - \alpha = 0$$
$$\mathcal{H}(\omega) = -\frac{1}{2i\omega}\left[(\alpha i\omega + \alpha - 1) + \sqrt{(\alpha i\omega + \alpha - 1)^2 + 4i\omega\alpha}\right] \tag{92}$$

Now, evaluating this expression at $i\omega = -s - i\epsilon$ gives

$$\mathcal{H}(is - \epsilon) = \frac{1}{2(s + i\epsilon)}\left[(-\alpha s - i\alpha\epsilon + \alpha - 1) + \sqrt{(-\alpha s - \alpha i\epsilon + \alpha - 1)^2 - 4(s + i\epsilon)\alpha}\right] \tag{93}$$

The radical has an imaginary solution in the $\epsilon \to 0$ limit provided that

$$s \in [s_-, s_+] \ , \ s_\pm = \left(1 \pm \frac{1}{\sqrt{\alpha}}\right)^2 \tag{94}$$

In this interval $[s_-, s_+]$, the density $\rho(s) = \lim_{\epsilon \to 0} \frac{1}{\pi}\text{Im}\mathcal{H}(is - \epsilon)$ has the form

$$\rho(s) = \frac{\alpha\sqrt{(s - s_-)(s - s_+)}}{2\pi s} \ , \ s \in [s_-, s_+] \tag{95}$$

which is precisely the bulk of the Marchenko-Pastur law.

G. Non-Proportional (Dimension-Free) Limit

We can imagine a situation where the original features are already infinite dimensional ($M \to \infty$ is taken first). This would correspond more naturally to the connection between infinite dimensional RKHS's induced by neural networks at infinite width (Bordelon et al., 2020; Canatar et al., 2021; Cheng & Montanari, 2022). Further, we will assume a trace class kernel $K(\boldsymbol{x}, \boldsymbol{x}') = \boldsymbol{\psi}(\boldsymbol{x}) \cdot \boldsymbol{\psi}(\boldsymbol{x}')$ for the base features $\boldsymbol{\psi}$ which diagonalizes over the data distribution $p(\boldsymbol{x})$ as

$$\int K(\boldsymbol{x}, \boldsymbol{x}')\phi_k(\boldsymbol{x}')p(\boldsymbol{x}')d\boldsymbol{x}' = \lambda_k\phi_k(\boldsymbol{x}) \quad , \quad \sum_{k=1}^{\infty}\lambda_k < \infty. \tag{96}$$

As before, we are concerned with the test and train losses

$$\mathcal{L}(t) = \sum_{k=1}^{\infty}\lambda_k v_k^0(t)^2 \ , \ \hat{\mathcal{L}}(t) = \frac{1}{P}\sum_{\mu=1}^{P} v_\mu^1(t)^2. \tag{97}$$

The appropriate scaling of our four fields of interest in this setting are

$$\boldsymbol{v}^1(t) = \boldsymbol{\Psi}\boldsymbol{v}^0(t) \ , \ \boldsymbol{v}^2(t) = \frac{1}{P}\boldsymbol{\Psi}^\top\boldsymbol{v}^1(t)$$
$$\boldsymbol{v}^3(t) = \boldsymbol{A}\boldsymbol{v}^2(t) \ , \ \boldsymbol{v}^4(t) = \frac{1}{N}\boldsymbol{A}^\top\boldsymbol{v}^3(t). \tag{98}$$

Following the cavity argument given in the previous section, we can approximate the the correlation and response functions as concentrating to arrive at the following field description of the training dynamics

$$\begin{aligned}
&\partial_t v_k^0(t) = -v_k^4(t) \\
&v^1(t) = u^1(t) + \frac{1}{P}\int ds\ R_{0,2}(t,s)v^1(s) \ , \ u^1(t) \sim \mathcal{GP}(0, C_0) \\
&v_k^2(t) = u_k^2(t) + \lambda_k \int ds\ R_1(t,s)v_k^0(s) \ , \ u_k^2(t) \sim \mathcal{GP}\left(0, \frac{1}{P}\lambda_k C^1\right) \\
&v^3(t) = u^3(t) + \frac{1}{N}\int ds\ R_{2,4}(t,s)v^3(s) \ , \ u^3(t) \sim \mathcal{GP}(0, C_2) \\
&v_k^4(t) = u_k^4(t) + \int ds\ R_3(t,s)v_k^2(s) \ , \ u_k^4(t) \sim \mathcal{N}\left(0, \frac{1}{N}C_3\right).
\end{aligned} \tag{99}$$

which are exactly the same equations as in the proportional limit except with the substitution $\nu \to N$ and $\alpha \to P$. The correlation and response functions have the form

$$\begin{aligned} C_0(t,s) &= \sum_{k=1}^{\infty} \lambda_k \left\langle v_k^0(t) v_k^0(s) \right\rangle \ , \ C_1(t,s) = \left\langle v^1(t) v^1(s) \right\rangle \\ C_2(t,s) &= \sum_{k=1}^{\infty} v_k^2(t) v_k^2(s) \ , \ C_3(t,s) = \left\langle v^3(t) v^3(s) \right\rangle \\ R_{0,2}(t,s) &= \sum_{k=1}^{\infty} \lambda_k \left\langle \frac{\delta v_k^0(t)}{\delta u_k^2(s)} \right\rangle \ , \ R_1(t,s) = \left\langle \frac{\delta v^1(t)}{\delta u^1(s)} \right\rangle \\ R_{2,4}(t,s) &= \sum_{k=1}^{\infty} \left\langle \frac{\delta v_k^2(t)}{\delta u_k^4(s)} \right\rangle \ , \ R_3(t,s) = \left\langle \frac{\delta v^3(t)}{\delta u^3(s)} \right\rangle \end{aligned} \tag{100}$$

which will all be $\mathcal{O}(1)$ under this scaling.

H. Effect of Ensembling and Bagging on Dynamics

H.1. What Does/Doesn't Concentrate in the DMFT Limit?

To help gain insight into bias and variance decompositions, we first provide a short primer on which entities concentrate over random draws of matrices $\boldsymbol{A}$ and $\boldsymbol{\Psi}$. For any distinct randomly sampled system, the following objects will always be the same in the asymptotic limit

1. The response functions $\{R_{0,2}(t,s), R_1(t,s), R_{2,4}(t,s), R_3(t,s)\}$
2. The correlation functions $\{C_i(t,s)\}_{i \in \{1,2,3,4\}}$.
3. The train and test loss dynamics

While the above quantities behave as concentrating or "self-averaging" random variables, many important quantities are not the same across different realizations of $\{\boldsymbol{A}, \boldsymbol{\Psi}\}$. For example,

1. The (random) entries of the vectors $\{\boldsymbol{v}^0(t), \boldsymbol{v}^1(t), \boldsymbol{v}^2(t), \boldsymbol{v}^3(t), \boldsymbol{v}^4(t)\}$.
2. The Gaussian sources $\{u^1(t), u^2(t), u^3(t), u^4(t)\}$ which appear in the large system size limit.

In particular, the first implies that the model outputs $f(\boldsymbol{x})$ will generally depend on random variations across datasets or model initializations. This means that we can consider drawing multiple realizations of, for example, projection matrices $\{\boldsymbol{A}_e\}_{e=1}^E$ and then training E separate models using each of them. Averaging these vectors gives us

$$\bar{\boldsymbol{v}}^0(t) = \frac{1}{E} \sum_{e=1}^{E} \boldsymbol{v}_e^0(t) \tag{101}$$

This operation will intuitively "average out" noise from the random projection matrices $\boldsymbol{A}_e$ and in the limit of infinite ensembling $E \to \infty$ will completely eliminate it.

H.2. Definition of Bias and Variance

We adopt the language of the fine-grained bias-variance decomposition in (Adlam & Pennington, 2020b). There, a given learned function generally depends on both the dataset $\mathcal{D}$ and initialization seed θ_0. We write this as $f_{\mathcal{D},\theta_0}$. The role of random initialization is played by the $\boldsymbol{A}$ matrix in our setting. For a given function, its variance over datasets and its *variance* over initializations are respectively given by

$$\mathrm{Var}_{\mathcal{D}} f \equiv \mathbb{E}_{\mathcal{D}} (f_{\mathcal{D},\theta_0} - \mathbb{E}_{\mathcal{D}}[f_{\mathcal{D},\theta_0}])^2 \tag{102}$$

$$\mathrm{Var}_{\theta_0} f \equiv \mathbb{E}_{\theta_0} (f_{\mathcal{D},\theta_0} - \mathbb{E}_{\theta_0}[f_{\mathcal{D},\theta_0}])^2 \tag{103}$$

Here $\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D},\theta_0}]$ and $\mathbb{E}_{\theta_0}[f_{\mathcal{D},\theta_0}]$ can be viewed as *infinitely bagged* or *infinitely ensembled* predictors respectively. The *bias* of a function over datasets or initializations is given by the test error of $\mathbb{E}_{\mathcal{D}}[f_{\mathcal{D},\theta_0}]$, $\mathbb{E}_{\theta_0}[f_{\mathcal{D},\theta_0}]$ respectively. The irreducible bias is given by $\mathbb{E}_{\mathcal{D},\theta_0}[f_{\mathcal{D},\theta_0}]$.