

H.3. Derivation

In this section, we consider the effect of ensembling over E random initial conditions and bagging over B random datasets. We let $\mathbf{v}_{e,b}^0(t)$ represent the weight discrepancy for model e on dataset b . Here e runs from 1 to E and b runs from 1 to B . The (e, b) th vector has dynamics:

$$\frac{d}{dt} \mathbf{v}_{e,b}^0(t) = - \left(\frac{1}{N} \mathbf{A}_e^\top \mathbf{A}_e \right) \left(\frac{1}{N} \mathbf{\Psi}_b^\top \mathbf{\Psi}_b \right) \mathbf{v}_{e,b}^0(t). \quad (104)$$

Ensembling and bagging would correspond to averaging these \mathbf{v}^0 's over these EB systems

$$\bar{\mathbf{v}}^0(t) = \frac{1}{EB} \sum_{e=1}^E \sum_{b=1}^B \mathbf{v}_{e,b}^0(t). \quad (105)$$

The key vectors to track for this computation are

$$\begin{aligned} \mathbf{v}_{e,b}^1(t) &= \frac{1}{\sqrt{M}} \mathbf{\Psi}_b \mathbf{v}_{e,b}^0(t) + \sigma \epsilon_b, \quad \mathbf{v}_{e,b}^2(t) = \frac{1}{\alpha \sqrt{M}} \mathbf{\Psi}_b^\top \mathbf{v}_{e,b}^1(t) \\ \mathbf{v}_{e,b}^3(t) &= \frac{1}{\sqrt{M}} \mathbf{A}_e \mathbf{v}_{e,b}^2(t), \quad \mathbf{v}_{e,b}^4(t) = \frac{1}{\nu \sqrt{M}} \mathbf{A}_e^\top \mathbf{v}_{e,b}^3(t). \end{aligned} \quad (106)$$

We can further show that the $\mathbf{v}_{e,b}^0$ and $\mathbf{v}_{e',b'}^0$ have response functions that decouple across e, b . Intuitively, giving the dynamical system e, b a kick should not alter the trajectory of the separate e', b' dynamical system, even if they share disorder $\{\mathbf{\Psi}, \mathbf{A}\}$. The DMFT description of the proportional limit yields the following integral equations for the v fields:

$$\begin{aligned} \partial_t v_{e,b,k}^0(t) &= -v_{e,b,k}^4(t) \\ v_{e,b}^1(t) &= u_{e,b}^1(t) + \frac{1}{\alpha} \int ds R_{0,2}(t,s) v_{e,b}^1(s) \\ v_{e,b,k}^2(t) &= u_{e,b,k}^2(t) + \lambda_k \int ds R_1(t,s) v_{e,b,k}^0(s) \\ v_{e,b}^3(t) &= u_{e,b}^3(t) + \frac{1}{\nu} \int ds R_{2,4}(t,s) v_{e,b}^1(s) \\ v_{e,b,k}^4(t) &= u_{e,b,k}^4(t) + \int ds R_3(t,s) v_{e,b,k}^2(s). \end{aligned} \quad (107)$$

Here, the response functions R are to be computed within a single system. In what follows, we will use $\langle \cdot \rangle$ to denote averages over the disorder, and explicitly write out any averages over the ensemble members and datasets.

The Gaussian variables in the DMFT have the following covariance

$$\begin{aligned} \langle u_{e,b}^1(t) u_{e',b'}^1(s) \rangle &= \delta_{b,b'} C_{e,e'}^0(t,s) \\ \langle u_{e,b,k}^2(t) u_{e',b',k}^2(s) \rangle &= \delta_{b,b'} \frac{\lambda_k}{\alpha} C_{1,e,e'}(t,s) \\ \langle u_{e,b}^3(t) u_{e',b'}^3(s) \rangle &= \delta_{e,e'} C_{2,b,b'}(t,s) \\ \langle u_{e,b,k}^4(t) u_{e',b',k}^4(s) \rangle &= \delta_{e,e'} \frac{1}{\nu} C_{3,b,b'}(t,s) \end{aligned} \quad (108)$$

The covariances above $C_{0,e,e'}, C_{1,e,e'}, C_{2,b,b'}, C_{3,b,b'}$ allow for different ensemble or dataset index but not both. We will use C_0, C_1, C_2, C_3 etc to represent the correlation functions *within a single system*. For instance, $C_{e,e'}^0(t,s) = \frac{1}{M} \sum_k \lambda_k \langle v_{e,b}^0(t) v_{e',b}^0(s) \rangle$ while $C^0 = \frac{1}{M} \sum_k \lambda_k \langle v_{e,b}^0(t) v_{e,b}^0(s) \rangle$. The correlation function of interest is thus

$$\begin{aligned} \mathcal{C}_{0,k,e,e'}(\omega, \omega') &= \mathcal{H}_k(\omega) \mathcal{H}_k(\omega') \left[(w_k^*)^2 + \frac{1}{\nu} \delta_{e,e'} \mathcal{C}_3(\omega, \omega') + \frac{\lambda_k}{\alpha} \mathcal{C}_{1,e,e'}(\omega, \omega') \right] \\ \mathcal{C}_{1,e,e'}(\omega, \omega') &= \mathcal{R}_1(\omega) \mathcal{R}_1(\omega') \mathcal{C}_{e,e'}^0(\omega, \omega') \\ \mathcal{C}_{2,b,b',k}(\omega, \omega') &= (i\omega)(i\omega') \mathcal{H}_k(\omega) \mathcal{H}_k(\omega') \frac{\lambda_k}{\alpha} \delta_{b,b'} \mathcal{C}_1(\omega, \omega') + \lambda_k^2 \mathcal{R}_1(\omega) \mathcal{R}_1(\omega') \mathcal{H}_k(\omega) \mathcal{H}_k(\omega') \left[(w_k^*)^2 + \frac{1}{\nu} \mathcal{C}_{3,b,b'}(\omega, \omega') \right] \\ \mathcal{C}_{3,b,b'}(\omega, \omega') &= \mathcal{R}_3(\omega) \mathcal{R}_3(\omega) \mathcal{C}_{2,b,b'}(\omega, \omega') \end{aligned} \quad (109)$$

We can combine the first two equations and the second two equations to identify the structure of the cross-ensemble and cross-dataset (across-system) correlations in terms of the marginal (within-system) correlation statistics

$$\begin{aligned}\mathcal{C}_{0,e,e'}(\omega, \omega') &= \frac{1}{1 - \gamma_0(\omega, \omega')} \frac{1}{M} \sum_k \lambda_k \mathcal{H}_k(\omega) \mathcal{H}_k(\omega') \left[(w_k^*)^2 + \frac{1}{\nu} \delta_{e,e'} \mathcal{C}_3(\omega, \omega') \right] \\ \gamma_0(\omega, \omega') &= \frac{1}{\alpha M} \sum_k \lambda_k^2 \mathcal{H}_k(\omega) \mathcal{H}_k(\omega') \mathcal{R}_1(\omega) \mathcal{R}_1(\omega') \mathcal{R}_3(\omega) \mathcal{R}_3(\omega') \\ \mathcal{C}_{2,b,b'}(\omega, \omega') &= \frac{1}{1 - \gamma_2(\omega, \omega')} \frac{1}{M} \sum_k \lambda_k \mathcal{H}_k(\omega) \mathcal{H}_k(\omega') \left[\mathcal{R}_1(\omega) \mathcal{R}_1(\omega') \lambda_k (w_k^*)^2 + \frac{1}{\alpha} (i\omega)(i\omega') \delta_{b,b'} \mathcal{C}_1(\omega, \omega') \right] \\ \gamma_2(\omega, \omega') &= \frac{1}{\nu M} \sum_k \lambda_k^2 \mathcal{H}_k(\omega) \mathcal{H}_k(\omega') \mathcal{R}_1(\omega) \mathcal{R}_1(\omega') \mathcal{R}_3(\omega) \mathcal{R}_3(\omega')\end{aligned}\quad (110)$$

These equations give the necessary cross-ensemble and cross-dataset correlations. Now we can consider the effect of ensembling and bagging on the dynamics. To do so, consider the Fourier transform of the bagged-ensembled error $\bar{v}_k^0(t) = \frac{1}{EB} \sum_{eb} v_{k,e,b}^0(t)$, which has the Fourier transform

$$\bar{v}_k^0(\omega) = \mathcal{H}_k(\omega) \left[w_k^* - \frac{1}{EB} \sum_{e,b} (u_{e,b,k}^4(\omega) + \mathcal{R}_3(\omega) u_{e,b,k}^2(\omega)) \right] \quad (111)$$

Computing the correlation function for this bagged-ensembled field random variable, we find

$$\begin{aligned}\langle \bar{v}_k^0(\omega) \bar{v}_k^0(\omega') \rangle &= \mathcal{H}_k(\omega) \mathcal{H}_k(\omega') \left[(w_k^*)^2 + \frac{1}{\nu E^2 B^2} \sum_{e,e',b,b'} \delta_{e,e'} \mathcal{C}_{3,b,b'}(\omega, \omega') + \frac{\lambda_k \mathcal{R}_3(\omega) \mathcal{R}_3(\omega')}{\alpha E^2 B^2} \sum_{ee'bb'} \delta_{b,b'} \mathcal{C}_{1,e,e'}(\omega, \omega') \right] \\ &= \mathcal{H}_k(\omega) \mathcal{H}_k(\omega') \left[(w_k^*)^2 + \frac{1}{\nu E B^2} \sum_{b,b'} \mathcal{C}_{3,b,b'}(\omega, \omega') + \frac{\lambda_k \mathcal{R}_3(\omega) \mathcal{R}_3(\omega')}{\alpha E^2 B^2} \sum_{ee'} \mathcal{C}_{1,e,e'}(\omega, \omega') \right] \\ &= \mathcal{H}_k(\omega) \mathcal{H}_k(\omega') (w_k^*)^2 \\ &\quad + \frac{1}{\nu E} \frac{\mathcal{H}_k(\omega) \mathcal{H}_k(\omega') \mathcal{R}_3(\omega) \mathcal{R}_3(\omega') \mathcal{R}_1(\omega) \mathcal{R}_1(\omega')}{1 - \gamma_2(\omega, \omega')} \left[\frac{1}{M} \sum_{\ell} \lambda_{\ell}^2 (w_{\ell}^*)^2 \mathcal{H}_{\ell}(\omega) \mathcal{H}_{\ell}(\omega') \right] \\ &\quad + \frac{1}{\nu \alpha E B} \frac{\mathcal{H}_k(\omega) \mathcal{H}_k(\omega') \mathcal{R}_3(\omega) \mathcal{R}_3(\omega') \mathcal{C}_1(\omega, \omega') (i\omega)(i\omega')}{1 - \gamma_2(\omega, \omega')} \left[\frac{1}{M} \sum_{\ell} \lambda_{\ell} \mathcal{H}_{\ell}(\omega) \mathcal{H}_{\ell}(\omega') \right] \\ &\quad + \frac{\lambda_k}{\alpha B} \frac{\mathcal{H}_k(\omega) \mathcal{H}_k(\omega') \mathcal{R}_1(\omega) \mathcal{R}_1(\omega') \mathcal{R}_3(\omega) \mathcal{R}_3(\omega')}{1 - \gamma_0(\omega, \omega')} \left[\frac{1}{M} \sum_{\ell} \lambda_{\ell} (w_{\ell}^*)^2 \mathcal{H}_{\ell}(\omega) \mathcal{H}_{\ell}(\omega') \right] \\ &\quad + \frac{\lambda_k}{\alpha \nu E B} \frac{\mathcal{H}_k(\omega) \mathcal{H}_k(\omega') \mathcal{R}_1(\omega) \mathcal{R}_1(\omega') \mathcal{R}_3(\omega) \mathcal{R}_3(\omega') \mathcal{C}_3(\omega, \omega')}{1 - \gamma_0(\omega, \omega')} \left[\frac{1}{M} \sum_{\ell} \lambda_{\ell} \mathcal{H}_{\ell}(\omega) \mathcal{H}_{\ell}(\omega') \right]\end{aligned}\quad (112)$$

The first term is the irreducible bias for mode k which is the loss for mode k when the learned function is averaged over all possible datasets and all possible projections. We see that the second term scales as $\frac{1}{\nu E}$ which will persist even if $B \alpha \rightarrow \infty$. Similarly, there is a term that is order $\frac{1}{\alpha B}$ which will persist even if $\nu E \rightarrow \infty$. Lastly, there are two terms which depend on both B, E . This is similar to the variance that is explained by the interaction of the dataset and the random projection (Adlam & Pennington, 2020b). The test loss is then a Fourier transform of the above function

$$\bar{\mathcal{L}}(t) = \frac{1}{M} \sum_k \lambda_k \langle \bar{v}_k^0(t)^2 \rangle. \quad (113)$$

If $E, B \rightarrow \infty$, then we obtain the stated *irreducible bias* of the main paper

$$\lim_{E,B \rightarrow \infty} \bar{\mathcal{L}}(t) = \frac{1}{M} \sum_k \lambda_k (w_k^*)^2 H_k(t)^2. \quad (114)$$

This is the error of the mean output function over all possible datasets and random projections of a certain size.

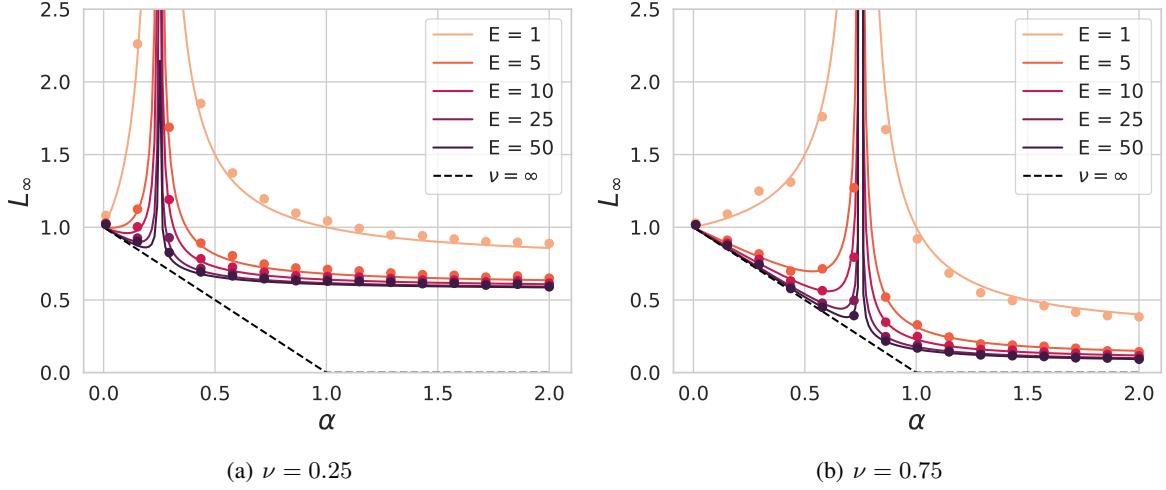


Figure 9. The infinite time limit of the loss when ensembling with isotropic features $\lambda_k = 1$ recovers prior results on ensembling and double descent (d’Ascoli et al., 2020; Adlam & Pennington, 2020b). There is an overfitting peak (double descent) at $\alpha = \nu$. In the *overparameterized regime* where $\alpha < \nu$, the infinite ensembled model matches the performance of the $\nu \rightarrow \infty$ limit. This is because the bias is limited by dataset size rather than model size. In the *underparameterized regime* $\alpha > \nu$, the infinite ensembled model *does not* achieve the loss of the infinite model due to a bias limited by ν .

H.4. Ensembling is Not Always Compute Optimal

For a compute budget $C = NET$, we find that ensembling does not provide as much benefit as increasing the size of the model. From the results in the last section, we note that ensembling reduces the variance. For this section, we consider the $P \rightarrow \infty$ limit. We let $\mathcal{B}(N, t)$ represent the bias and $\mathcal{V}(N, t)$ represent the variance within a single ensemble. The loss at fixed compute then takes the form

$$\mathcal{L}(\nu, C, t) = \mathcal{B}(\nu, t) + \frac{1}{\nu E} \mathcal{V}(\nu, t). \quad (115)$$

For any ν which satisfies the condition that

$$\frac{\partial}{\partial \nu} \mathcal{B}(\nu, t) \leq 0, \quad \frac{\partial}{\partial \nu} \mathcal{V}(\nu, t) \leq 0 \quad (116)$$

we have that ensembling is strictly dominated by increasing ν .

I. White Bandlimited Model

To gain intuition for the model, we can first analyze the case where $\lambda_k = 1$, which has a simpler DMFT description since each of the M features are statistically identical. We illustrate the dependence of the loss on model size ν and training time t for $\alpha < 1$ in Figure 10. We note that the loss can be non-monotonic in ν at late training times, but that monotonicity is maintained for optimal early stopping, similar to results on optimal regularization in linear models (Advani et al., 2020) and random feature models (Mei & Montanari, 2022; Simon et al., 2023).

I.1. Derivation

In the case of all $\lambda_k = 1$ we have the following definitions

$$\begin{aligned} \mathcal{R}_1(\omega) &= 1 - \frac{1}{\alpha} \frac{\mathcal{R}_1(\omega)\mathcal{R}_3(\omega)}{i\omega + \mathcal{R}_1(\omega)\mathcal{R}_3(\omega)} \\ \mathcal{R}_3(\omega) &= 1 - \frac{1}{\nu} \frac{\mathcal{R}_1(\omega)\mathcal{R}_3(\omega)}{i\omega + \mathcal{R}_1(\omega)\mathcal{R}_3(\omega)} \end{aligned} \quad (117)$$