A. Derivation of Dynamical Model of Scaling Laws

We investigate the simplest possible model which can exhibit task-dependent time, model size and finite data bottlenecks. We therefore choose to study a linear model with projected features

$$f(\boldsymbol{x}) = \frac{1}{\sqrt{N}} \boldsymbol{w}^\top \left(\frac{1}{\sqrt{M}} \boldsymbol{A} \boldsymbol{\psi}(\boldsymbol{x}) \right) , \ y(\boldsymbol{x}) = \frac{1}{\sqrt{M}} \boldsymbol{w}_\star \cdot \boldsymbol{\psi}(\boldsymbol{x}). \tag{21}$$

The weights $\boldsymbol{w}$ are updated with gradient descent on a random training dataset which has (possibly) noise corrupted target values $y_\mu = y(\boldsymbol{x}_\mu) + \sigma\epsilon_\mu$. This leads to the following gradient flow dynamics

$$\frac{\partial}{\partial t}\boldsymbol{w}(t) = \frac{\sqrt{M}}{P\sqrt{N}}\sum_{\mu=1}^{P}(y_\mu - f_\mu)\boldsymbol{A}\boldsymbol{\psi}_\mu = \frac{1}{\sqrt{N}}\boldsymbol{A}\left(\frac{1}{P}\sum_{\mu=1}^{P}\boldsymbol{\psi}_\mu\left[\boldsymbol{\psi}_\mu^\top\left(\boldsymbol{w}_\star - \frac{1}{\sqrt{N}}\boldsymbol{A}^\top\boldsymbol{w}\right) + \sqrt{M}\sigma\epsilon_\mu\right]\right). \tag{22}$$

We introduce the variable $\boldsymbol{v}^0 = \boldsymbol{w}_\star - \frac{1}{\sqrt{N}}\boldsymbol{A}^\top\boldsymbol{w}$ to represent the residual error of the learned weight vector. This residual error has the following dynamics:

$$\partial_t \boldsymbol{v}^0(t) = -\left(\frac{1}{N}\boldsymbol{A}^\top\boldsymbol{A}\right)\left[\left(\frac{1}{P}\boldsymbol{\Psi}^\top\boldsymbol{\Psi}\right)\boldsymbol{v}^0(t) + \frac{\sigma}{\alpha\sqrt{M}}\boldsymbol{\Psi}^\top\boldsymbol{\epsilon}\right]. \tag{23}$$

The entries of each matrix are treated as random with $\Psi_k^\mu \sim \mathcal{N}(0, \lambda_k)$ and $A_{jk} \sim \mathcal{N}(0,1)$. To study the dynamical evolution of the test error $\mathcal{L}(t) = \frac{1}{M}\boldsymbol{v}^0(t)^\top\boldsymbol{\Lambda}\boldsymbol{v}^0(t) + \sigma^2$, we introduce the sequence of vectors

$$\boldsymbol{v}^1(t) = \frac{1}{\sqrt{M}}\boldsymbol{\Psi}\boldsymbol{v}^0(t) + \sigma\boldsymbol{\epsilon} , \ \boldsymbol{v}^2(t) = \frac{1}{\alpha\sqrt{M}}\boldsymbol{\Psi}^\top\boldsymbol{v}^1(t)$$

$$\boldsymbol{v}^3(t) = \frac{1}{\sqrt{M}}\boldsymbol{A}\boldsymbol{v}^2(t) , \ \boldsymbol{v}^4(t) = \frac{1}{\nu\sqrt{M}}\boldsymbol{A}^\top\boldsymbol{v}^3(t). \tag{24}$$

The train and test losses can be computed from the $\boldsymbol{v}^0$ and $\boldsymbol{v}^1$ fields

$$\hat{\mathcal{L}}(t) = \frac{1}{P}\sum_{\mu=1}^{P}v_\mu^1(t)^2 , \ \mathcal{L}(t) = \frac{1}{M}\sum_{k=1}^{M}\lambda_k v_k^0(t)^2 + \sigma^2. \tag{25}$$

In the next section, we derive a statistical description of the dynamics in an appropriate asymptotic limit using dynamical mean field theory methods.

A.1. DMFT Equations for the Asymptotic Limit

Standard field theoretic arguments such as the cavity or path integral methods can be used to compute the effective statistical description of the dynamics in the limit of large M, N, P with fixed ratios $\alpha = P/M$ and $\nu = \frac{N}{M}$ (see Appendix B). This computation gives us the following statistical description of the dynamics.

$$v^1(t) = u^1(t) + \frac{1}{\alpha}\int ds R_{0,2}(t,s)v_1(s) + \sigma\epsilon , \ u^1(t) \sim \mathcal{GP}(0, C_0) , \ \epsilon \sim \mathcal{N}(0,1),$$

$$v_k^2(t) = u_k^2(t) + \lambda_k \int ds R_1(t,s)v_k^0(s) , \ u_k^2(t) \sim \mathcal{GP}\left(0, \frac{1}{\alpha}\lambda_k C_1\right),$$

$$v^3(t) = u^3(t) + \frac{1}{\nu}\int ds R_{2,4}(t,s)v^3(s) , \ u^3(t) \sim \mathcal{GP}(0, C_2), \tag{26}$$

$$v_k^4(t) = u_k^4(t) + \int ds R_3(t,s)v_k^2(s) , \ u_k^4(t) \sim \mathcal{GP}\left(0, \frac{1}{\nu}C_3\right),$$

$$\partial_t v_k^0(t) = -v_k^4(t).$$

The correlation and response functions obey

$$C_0(t,s) = \frac{1}{M}\sum_k \lambda_k \left\langle v_k^0(t)v_k^0(s)\right\rangle \;,\; C_1(t,s) = \left\langle v^1(t)v^1(s)\right\rangle \;,\; C_2(t,s) = \frac{1}{M}\sum_{k=1}^M \left\langle v_k^2(t)v_k^2(s)\right\rangle$$

$$R_{0,2}(t,s) = \frac{1}{M}\sum_k \lambda_k \left\langle \frac{\delta v_k^0(t)}{\delta u_k^2(s)}\right\rangle \;,\; R_{2,4}(t,s) = \frac{1}{M}\sum_k \left\langle \frac{\delta v_k^2(t)}{\delta u_k^4(s)}\right\rangle$$

$$R_1(t,s) = \left\langle \frac{\delta v^1(t)}{\delta u^1(s)}\right\rangle \;,\; R_3(t,s) = \left\langle \frac{\delta v^3(t)}{\delta u^3(s)}\right\rangle$$

These equations are exact in the joint proportional limit for any value of α, ν.

A.2. Closing the Equations for the Order Parameters

Though we expressed the dynamics in terms of random fields, we stress in this section that all of the dynamics for the correlation and response functions close in terms of integro-differential equations. To shorten the expression, we will provide the expression for $\beta = 0$, but momentum can easily be added back by making the substitution $\partial_t \to \partial_t + \beta\partial_t^2$.

First, our closed integral equations for the response functions are

$$R_{0,2,k}(t,s) = -\int dt'\Theta(t-t')R_3(t',s) - \lambda_k \int dt'dt''dt'''\Theta(t-t')R_3(t',t'')R_1(t'',t''')R_{0,2,k}(t''',s)$$

$$R_1(t,s) = \delta(t-s) + \frac{1}{\alpha}\int dt' R_{0,2}(t,t')R_1(t',s)$$

$$R_{2,4,k}(t,s) = -\lambda_k \int dt'dt'' R_1(t,t')\Theta(t'-t'') - \lambda_k \int dt'dt''dt''' R_1(t,t')\Theta(t'-t'')R_3(t'',t''')R_{2,4,k}(t''',s)$$

$$R_3(t,s) = \delta(t-s) + \frac{1}{\nu}\int dt' R_{2,4}(t,t')R_3(t',s)$$

$$R_{0,2}(t,s) = \frac{1}{M}\sum_k \lambda_k R_{0,2,k}(t,s) \;,\; R_{2,4}(t,s) = \frac{1}{M}\sum_k R_{2,4,k}(t,s) \tag{27}$$

We note that these equations imply causality in all of the response functions since $R(t,s) = 0$ for $t < s$. Once these equations are solved for the response functions, we can determine the correlation functions, which satisfy

$$\partial_{ts}^2 C_{0,k}(t,s) = -\lambda_k \int dt'dt'' R_3(t,t')R_1(t',t'')\partial_s C_{0,k}(t'',s)$$

$$- \lambda_k \int ds' R_3(s,s')R_1(s',s'')\partial_t C_{0,k}(t,s'')$$

$$+ \lambda_k^2 \int dt'dt''ds'ds'' R_3(t,t')R_1(t',t'')R_3(s,s')R_1(s',s'')C_{0,k}(t'',s'')$$

$$- (w_k^\star)^2\delta(t)\delta(s) - \frac{1}{\nu}C_3(t,s) - \frac{1}{\alpha}\int dt'ds' R_3(t,t')R_3(s,s')C_1(t',s')$$

$$C_1(t,s) = \int dt' R_1(t,t')R_1(s,s')C_0(t',s')$$

$$C_{2,k}(t,s) = -\lambda_k \int dt'dt'' R_1(t,t')R_3(t',t'')C_{2,k}(t'',s) - \lambda_k \int ds'ds'' R_1(s,s')R_3(s',s'')C_2(t,s'')$$

$$+ \lambda_k^2 \int dt'dt''ds'ds'' R_1(t,t')R_3(t',t'')R_1(s,s')R_3(s',s'')C_{2,k}(t'',s'')$$

$$C_3(t,s) = \int dt'ds' R_3(t,t')R_3(s,s')C_2(t',s') \tag{28}$$

Solving these closed equations provide the complete statistical characterization of the limit. The test and train losses are given by the time-time diagonal of $C_0(t,t), C_1(t,t)$.

A.3. Time-translation Invariant (TTI) Solution to Response Functions

From the structure of the above equations, the response functions are time-translation invariant (TTI) since they are only functionals of TTI $\delta(t-s)$ Dirac-Delta function and $\Theta(t-s)$ Heaviside step-function. As a consequence, we write each of our response functions in terms of their Fourier transforms

$$R(t,s) = R(t-s) = \int_{-\infty}^{\infty} \frac{d\omega}{2\pi} e^{i\omega(t-s)} \mathcal{R}(\omega). \tag{29}$$

Using the fact that

$$\delta(\tau) = \int \frac{d\omega}{2\pi} e^{i\omega\tau} \ , \ \Theta(\tau) = \lim_{\epsilon \to 0^+} e^{-\epsilon\tau} \Theta(\tau) = \lim_{\epsilon \to 0^+} \int \frac{d\omega}{2\pi} \frac{e^{i\omega\tau}}{\epsilon + i\omega} \tag{30}$$

We will keep track of the regulator ϵ and consider $\epsilon \to 0^+$ at the end of the computation. The resulting DMFT equations for the response functions have the following form in Fourier space

$$\mathcal{R}_1(\omega) = 1 + \frac{1}{\alpha} \mathcal{R}_{2,4}(\omega) \mathcal{R}_1(\omega)$$

$$\mathcal{R}_3(\omega) = 1 + \frac{1}{\nu} \mathcal{R}_{2,4}(\omega) \mathcal{R}_3(\omega)$$

$$\mathcal{R}_{0,2}(\omega) = -\frac{1}{M} \sum_k \frac{\lambda_k}{\epsilon + i\omega + \lambda_k \mathcal{R}_1(\omega) \mathcal{R}_3(\omega)} \mathcal{R}_3(\omega)$$

$$\mathcal{R}_{2,4}(\omega) = -\frac{1}{M} \sum_k \frac{\lambda_k}{\epsilon + i\omega + \lambda_k \mathcal{R}_1(\omega) \mathcal{R}_3(\omega)} \mathcal{R}_1(\omega) \tag{31}$$

where $\epsilon \to 0$ will be taken after. Combining these equations, we arrive at the simple set of coupled equations

$$\mathcal{R}_1(\omega) = 1 - \frac{1}{P} \sum_k \frac{\lambda_k \mathcal{R}_3(\omega) \mathcal{R}_1(\omega)}{\epsilon + i\omega + \lambda_k \mathcal{R}_1(\omega) \mathcal{R}_3(\omega)}$$

$$\mathcal{R}_3(\omega) = 1 - \frac{1}{N} \sum_k \frac{\lambda_k \mathcal{R}_1(\omega) \mathcal{R}_3(\omega)}{\epsilon + i\omega + \lambda_k \mathcal{R}_1(\omega) \mathcal{R}_3(\omega)} \tag{32}$$

After solving these equations for all ω, we can invert the dynamics of $v_k^0(t)$ to obtain its Fourier transform

$$v_k^0(\omega) = \mathcal{H}_k(\omega) \left[w_k^\star - u_k^4(\omega) - \lambda_k \mathcal{R}_3(\omega) u_k^2(\omega) \right]$$

$$\mathcal{H}_k(\omega) \equiv \frac{1}{\epsilon + i\omega + \lambda_k \mathcal{R}_1(\omega) \mathcal{R}_3(\omega)} \tag{33}$$

where we defined the transfer functions $\mathcal{H}_k(\omega)$. From this equation, we can compute $v_k^0(t)$ through inverse Fourier-transformation and then compute the correlation function to calculate the test error. An interesting observation is that the response functions $\mathcal{R}_1(\omega), \mathcal{R}_3(\omega)$ alter the pole structure in the transfer function, generating ν, α dependent timescales of convergence.

A.4. Fourier Representations for Correlation Functions

While the response functions are TTI, the correlation functions transparently are not (if the time-time diagonal $C_0(t,t)$ did not evolve, then the loss $\mathcal{L}(t)$ wouldn't change!). We therefore define the need to define the double Fourier transform $\mathcal{C}(\omega, \omega')$ for each correlation function $C(t,s)$

$$\mathcal{C}(\omega, \omega') = \int dt\,ds\, e^{-i\omega t - i\omega' s} C(t,s) \ , \ C(t,s) = \int \frac{d\omega}{2\pi} \frac{d\omega'}{2\pi} e^{i\omega t + i\omega' s} \mathcal{C}(\omega, \omega') \tag{34}$$