Two-point deterministic equivalence for SGD in random feature models

Alexander Atanasov* ATANASOV@G.HARVARD.EDU
Harvard University, Cambridge, MA, USA

Blake Bordelon* BLAKE_BORDELON@G.HARVARD.EDU
Harvard University, Cambridge, MA, USA

Jacob A. Zavatone-Veth* JZAVATONEVETH@FAS.HARVARD.EDU
Harvard University, Cambridge, MA, USA

Courtney Paquette COURTNEY.PAQUETTE@MCGILL.CA
McGill University, Montréal, QC, CA

Cengiz Pehlevan CPEHLEVAN@SEAS.HARVARD.EDU
Harvard University, Cambridge, MA, USA

Abstract

We derive a novel deterministic equivalence for the two-point function of a random matrix resolvent. Using this result, we give a unified derivation of the performance of a wide variety of high-dimensional linear models trained with stochastic gradient descent. This includes high-dimensional linear regression, kernel regression, and random feature models. Our results include previously known asymptotics as well as novel ones.

1. Introduction

Modern deep learning practice is governed by the surprising predictability of performance improvement with increases in the scale of data, model size, and compute [15]. Often, the scaling of performance as a function of these quantities exhibits remarkably regular power law behavior, termed a neural scaling law [2, 6, 11–14, 16, 17, 20, 30]. Given the relatively universal behavior of the exponents across architectures and optimizers [10, 16, 17], one might hope that relatively simple models of information processing systems might be able to recover the same types of scaling laws.

The (stochastic) gradient descent (SGD) dynamics in random feature models were analyzed in recent works [8, 18, 24], exhibiting a surprising breadth of scaling behavior and captures several interesting phenomena in deep network training. Each of the above works has isolated various effects that can hurt performance compared to the idealized infinite data and infinite model size limits. The model was first studied in [8], where the bottlenecks due to finite width and finite dataset size were computed and, for certain data structure, resulted in a Chinchilla-type scaling result as in [16]. In [24], the effect of finite models size and online SGD noise was studied, and it was shown that under certain conditions these effects could lead to worse scaling exponents with the number of iterations than those one would naively calculate from a leading order picture.

* AA, BB, and JAZ-V contributed equally to this work.

In this work, we aim to unify these prior results by providing a novel deterministic equivalence result for correlations of resolvent matrices evaluated at an arbitrary pair of arguments. This result allows us to combine the interaction of limited data, limited features, and SGD noise to model the stochastic process induced by SGD in the random feature model. We leverage the connections between deterministic equivalence and free probability highlighted in [4]. Specifically, we use the properties of the S-transform and the diagrammatic expansion to obtain the asymptotic expressions for the train and test losses in time in the linear random feature model. Our results recover those obtained via a dynamical mean field theory (DMFT) approach by [8] and with those obtained using deterministic equivalence techniques in [24].

2. Setup

2.1. SGD in linear random feature models models

In this paper, we will study linear random feature models

$$f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{F}\boldsymbol{v}, \tag{1}$$

where $\boldsymbol{v}$ is trainable and $\boldsymbol{F}$ is fixed and random. Here, $\boldsymbol{x} \in \mathbb{R}^D$, $\boldsymbol{F} \in \mathbb{R}^{D\times N}$, $\boldsymbol{v} \in \mathbb{R}^N$. We also define the *effective learned weights* $\boldsymbol{w} \equiv \boldsymbol{F}\boldsymbol{v}$. We train $f(\boldsymbol{x})$ to fit a set of labels y generated from Gaussian covariates with covariance $\boldsymbol{\Sigma}$ by a linear teacher $\bar{\boldsymbol{w}}$, corrupted by noise $\epsilon_\mu \sim \mathcal{N}(0, \sigma^2)$:

$$y_\mu = \bar{\boldsymbol{w}} \cdot \boldsymbol{x}_\mu + \epsilon_\mu, \quad \boldsymbol{x}_\mu \sim \mathcal{N}(\boldsymbol{0}, \boldsymbol{\Sigma}). \tag{2}$$

We take $\boldsymbol{F}$ to have i.i.d. Gaussian entries with variance $1/N$. In what follows, we will drop the noise term, as it amounts to adding a vanishingly small eigenvalue λ_∞ to $\boldsymbol{\Sigma}$ with a corresponding target weight $\bar{w}_\infty^2 = \sigma^2/\lambda_\infty$ and taking $\lambda_\infty \to 0$.

Using stochastic gradient descent (SGD), we minimize the empirical risk $\hat{R}$ estimated from P examples as a proxy for the population risk R:

$$\hat{R} = \frac{1}{P}\sum_{\mu=1}^{P}(y_\mu - f(\boldsymbol{x}_\mu))^2, \quad R = \mathbb{E}_{\boldsymbol{x},y}(y - f(\boldsymbol{x}))^2. \tag{3}$$

That is, at each step t, we sample a random batch $\mathcal{B}_t$ of $B < P$ training points (with replacement) and update the weights in proportion to the loss gradient on that batch, with learning rate $\eta > 0$:

$$\boldsymbol{v}_{t+1} = \boldsymbol{v}_t - \eta \nabla_{\boldsymbol{v}} \hat{R}_{\mathcal{B}_t}. \tag{4}$$

As they will frequently appear in subsequent expressions, we define the design matrix $\boldsymbol{X} \in \mathbb{R}^{P\times D}$ and the vector of training labels $\boldsymbol{y} \in \mathbb{R}^P$. With this, we define the two empirical covariances:

$$\hat{\boldsymbol{\Sigma}} \equiv \frac{1}{P}\sum_{\mu=1}^{P}\boldsymbol{x}_\mu\boldsymbol{x}_\mu^\top = \frac{1}{P}\boldsymbol{X}^\top\boldsymbol{X}, \quad \hat{\boldsymbol{\Sigma}}_t \equiv \frac{1}{B}\sum_{\mu\in\mathcal{B}_t}\boldsymbol{x}_{\mu,t}\boldsymbol{x}_{\mu,t}^\top. \tag{5}$$

Note, $\mathbb{E}_{\mathcal{B}_t}\hat{\boldsymbol{\Sigma}}_t = \hat{\boldsymbol{\Sigma}}$ and $\mathbb{E}_{\boldsymbol{X}}\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}$. This highlights the two levels of randomness from the data: the randomness due to the choice of batch and the randomness due to the choice of the training set.

2.2. A reduced model for high-dimensional risk dynamics

We are interested in the asymptotic behavior of the empirical and population risks in the high-dimensional limit $D, N, P \to \infty$ with fixed ratios D/P and D/N. Moreover, we will focus on a limit in which we take the learning rate $\eta \to 0$ limit while keeping the **SGD temperature** $\chi = \eta/B$ constant [7]. In this limit, we will use a reduced model for the dynamics of the training and generalization error under SGD, following previous works that argue this reduced model should be asymptotically equivalent in high dimensions [7, 21–24]. Deferring the details of the derivation of this reduced model to Appendix A, the result is a pair of coupled Volterra integral equations:

$$\hat{R}(t) = \underbrace{\bar{\boldsymbol{w}}^\top e^{-2t\hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top}\hat{\boldsymbol{\Sigma}}\bar{\boldsymbol{w}}}_{\hat{\mathcal{F}}(t)} + \chi\int_0^t \underbrace{\mathrm{Tr}[e^{-2(t-s)\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}}(\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}})^2]}_{\hat{\mathcal{K}}(t-s)}\hat{R}_s ds, \tag{6}$$

$$R(t) = \underbrace{\bar{\boldsymbol{w}}^\top e^{-t\hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top}\boldsymbol{\Sigma}e^{-t\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}}\bar{\boldsymbol{w}}}_{\mathcal{F}(t)} + \chi\int_0^t \underbrace{\mathrm{Tr}[e^{-2(t-s)\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}}\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top\boldsymbol{\Sigma}]}_{\mathcal{K}(t-s)}\hat{R}_s ds. \tag{7}$$

For linear regression, $\boldsymbol{F}\boldsymbol{F}^\top$ is replaced by the identity matrix. In both equations, the first term on the right hand side is referred to as the **forcing term**, while the second term we will refer to as the response or **kernel** term. It is the second term that is due to SGD noise, and would go away in the limit of $\chi \to 0$. The kernel term consists of a convolution of the population risk with the train and test kernels $\hat{\mathcal{K}}, \mathcal{K}$ respectively. This notation is adopted from [24]. From these results, one can recover long-time limits that correspond to ordinary ridge regression using final value theorems (Appendix B).

2.3. Strong deterministic equivalence

The goal of our work is to derive **strong deterministic equivalents** for $\hat{R}_t$ and R_t. Here, we briefly recall the definition and standard example of strong deterministic equivalence, deferring a complete description to Appendix C. We say that two (possibly random) $N \times N$ matrices $\boldsymbol{A}$ and $\boldsymbol{B}$ are deterministically equivalent, and write $\boldsymbol{A} \simeq \boldsymbol{B}$, if $\mathrm{tr}(\boldsymbol{A}\boldsymbol{M})/\,\mathrm{tr}(\boldsymbol{B}\boldsymbol{M}) \to 1$ in probability as $N \to \infty$ for any sequence of bounded-spectral-norm test matrices $\boldsymbol{M}$. Here, $\mathrm{tr}(\cdot) = \mathrm{Tr}(\cdot)/N$ is the normalized trace. The classic example of such a strong deterministic equivalence is

$$\boldsymbol{A}\boldsymbol{B}(\lambda + \boldsymbol{A}\boldsymbol{B})^{-1} \simeq \boldsymbol{A}(\boldsymbol{A} + \kappa)^{-1} \tag{8}$$

for $\boldsymbol{A}$ fixed and $\boldsymbol{B}$ Wishart, where $\kappa = \lambda S_{\boldsymbol{B}}(\mathrm{df}_{\boldsymbol{A}}(\kappa))$ has the interpretation of a ridge parameter **renormalized** by the randomness in $\boldsymbol{B}$. In the self-consistent equation for κ, $S_{\boldsymbol{B}}$ is the S-transform of the matrix $\boldsymbol{B}$, and $\mathrm{df}_{\boldsymbol{A}}(\kappa) = \mathrm{tr}[\boldsymbol{A}(\boldsymbol{A} + \kappa)^{-1}]$ is the degrees of freedom. We define these quantities in more detail in Appendix C.

3. When is the one-point resolvent insufficient?

From the previous section, we see that in all of the settings of interest, the population forcing function can be represented as

$$\mathcal{F}(t) = \bar{\boldsymbol{w}}^\top e^{-t\boldsymbol{A}\boldsymbol{B}}\boldsymbol{M}e^{-t\boldsymbol{B}\boldsymbol{A}}\bar{\boldsymbol{w}}. \tag{9}$$

where $\boldsymbol{A}$ and $\boldsymbol{B}$ are random symmetric matrices and $\boldsymbol{M}$ is a fixed symmetric matrix. Under the conditions that $\{\boldsymbol{M}, \boldsymbol{A}, \boldsymbol{B}\}$ all jointly commute, one can write this expression in Fourier space as:

$$\mathcal{F}(t) = \bar{\boldsymbol{w}}^\top e^{-2t\boldsymbol{A}\boldsymbol{B}}\boldsymbol{M}\bar{\boldsymbol{w}} = \int_\omega e^{2i\omega t}\bar{\boldsymbol{w}}^\top\left(i\omega + \boldsymbol{A}\boldsymbol{B}\right)^{-1}\boldsymbol{M}\bar{\boldsymbol{w}}, \tag{10}$$

where we write $\int_\omega(\cdot) = \frac{1}{2\pi}\int(\cdot)\,d\omega$. Under this commutativity condition, computing sharp asymptotics at finite time for the above quantity amounts to computing a strong deterministic equivalent for the random matrix $(i\omega + \boldsymbol{A}\boldsymbol{B})^{-1}$. Such strong deterministic equivalences have been highlighted in a variety of recent literature [4, 5, 19]. However, under a variety of settings, the matrices $\{\boldsymbol{A}, \boldsymbol{B}, \boldsymbol{M}\}$ will not commute, necessitating a different approach. Two such settings of interest are as follows:

(1). When the dataset is finite, $\boldsymbol{A} = \hat{\boldsymbol{\Sigma}}$ and the population covariance $\boldsymbol{\Sigma}$ does not generally commute with $\boldsymbol{A}$, regardless of whether $\boldsymbol{B}$ is included. Thus, even in the case of linear regression at finite t, one requires two-point resolvents. (2). A random feature model at finite model size N and finite dataset size P will not have $\boldsymbol{A} = \boldsymbol{F}\boldsymbol{F}^\top$ and $\boldsymbol{B} = \hat{\boldsymbol{\Sigma}}$ commute. This is why the general finite N, P, t expressions of Bordelon et. al [8] required computing two point resolvents that are functions of two frequencies ω, ω'. At finite N but infinite P, $\boldsymbol{A} = \boldsymbol{M} = \boldsymbol{\Sigma}$ and one-point resolvents are in fact sufficient as in the work of Paquette et al [24]. To address these more general settings, we derive a novel set of "two-point" deterministic equivalences.

4. Two-point deterministic equivalents and risk asymptotics

If the relevant matrices do not commute, we write the forcing term in bifrequency form:

$$\mathcal{F}(t) = \int_{\omega,\omega'} e^{it(\omega+\omega')}\mathcal{F}(\omega,\omega') \tag{11}$$

$$\mathcal{F}(\omega,\omega') = \bar{\boldsymbol{w}}^\top(\hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top + i\omega)^{-1}\boldsymbol{\Sigma}(\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}} + i\omega')^{-1}\bar{\boldsymbol{w}}. \tag{12}$$

where we adopt the shorthand $\int_{\omega,\omega'}(\cdot) = \frac{1}{(2\pi)^2}\int(\cdot)\, d\omega\, d\omega'$. We must then derive a two-point deterministic equivalent for the resolvent $(\hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top + i\omega)^{-1}$. Abstractly, we want to derive a deterministic equivalent for

$$(\lambda + \boldsymbol{A}\boldsymbol{B})^{-1}\boldsymbol{M}(\lambda' + \boldsymbol{B}\boldsymbol{A})^{-1}. \tag{13}$$

where $\boldsymbol{A}$ and $\boldsymbol{M}$ are deterministic matrices, and $\boldsymbol{B}$ is an isotropic Wishart matrix and $\lambda, \lambda' \in \mathbb{C}$. In Appendix D, we show using a diagrammatic argument that

$$(\lambda + \boldsymbol{A}\boldsymbol{B})^{-1}\boldsymbol{M}(\lambda' + \boldsymbol{B}\boldsymbol{A})^{-1} \simeq S_{\boldsymbol{B}}S'_{\boldsymbol{B}}\left[\boldsymbol{G}_{\boldsymbol{A}}\boldsymbol{M}\boldsymbol{G}'_{\boldsymbol{A}} + \boldsymbol{G}_{\boldsymbol{A}}\boldsymbol{A}^2\boldsymbol{G}'_{\boldsymbol{A}}\frac{q\,\mathrm{tr}[\boldsymbol{G}_{\boldsymbol{A}}\boldsymbol{M}\boldsymbol{G}'_{\boldsymbol{A}}]}{1 - q\mathrm{df}_2(\kappa,\kappa')}\right]. \tag{14}$$

where

$$\begin{aligned}
S_{\boldsymbol{B}} &= S_{\boldsymbol{B}}(-\mathrm{df}^1_{\boldsymbol{AB}}(\lambda)), & S'_{\boldsymbol{B}} &= S_{\boldsymbol{B}}(-\mathrm{df}^1_{\boldsymbol{AB}}(\lambda')),\\
\kappa &= \lambda S_{\boldsymbol{B}}, & \kappa' &= \lambda' S'_{\boldsymbol{B}},\\
\boldsymbol{G}_{\boldsymbol{A}} &= (\kappa + \boldsymbol{A})^{-1}, & \boldsymbol{G}'_{\boldsymbol{A}} &= (\kappa' + \boldsymbol{A})^{-1},\\
g &= \mathrm{tr}[\boldsymbol{G}_{\boldsymbol{A}}], & g' &= \mathrm{tr}[\boldsymbol{G}'_{\boldsymbol{A}}],\\
\boldsymbol{T}_{\boldsymbol{A}} &= \boldsymbol{A}(\kappa + \boldsymbol{A})^{-1}, & \boldsymbol{T}'_{\boldsymbol{A}} &= \boldsymbol{A}(\kappa' + \boldsymbol{A})^{-1}\\
&& \mathrm{df}_2(\kappa,\kappa') &= \mathrm{tr}[\boldsymbol{A}^2\boldsymbol{G}_{\boldsymbol{A}}\boldsymbol{G}'_{\boldsymbol{A}}].
\end{aligned} \tag{15}$$

This is the main technical result of our work.

We can now state the main result of our work, which applies this two-point deterministic equivalent to obtain the asymptotic behavior of the random feature model risk. We derive this result in Appendix F, after a warm-up step of considering linear regression without random features in

Appendix E. Let

$$
\begin{aligned}
\mathrm{df}_1 \equiv \mathrm{df}^1_{\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}}(\omega), &\quad \mathrm{df}'_1 \equiv \mathrm{df}^1_{\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}}(\omega'),\\
S_{\boldsymbol{W}} \equiv \frac{1}{1-\frac{D}{P}\mathrm{df}_1}, &\quad S'_{\boldsymbol{W}} \equiv \frac{1}{1-\frac{D}{P}\mathrm{df}'_1},\\
S_{\boldsymbol{F}\boldsymbol{F}^\top} \equiv \frac{1}{1-\frac{D}{N}\mathrm{df}_1}, &\quad S'_{\boldsymbol{F}\boldsymbol{F}^\top} \equiv \frac{1}{1-\frac{D}{N}\mathrm{df}'_1},\\
S \equiv S_{\boldsymbol{W}}S_{\boldsymbol{F}\boldsymbol{F}^\top}, &\quad S' \equiv S'_{\boldsymbol{W}}S'_{\boldsymbol{F}\boldsymbol{F}^\top},\\
\omega_1 \equiv S_{\boldsymbol{W}}\omega, &\quad \omega'_1 \equiv S'_{\boldsymbol{W}}\omega',\\
\omega_2 \equiv S_{\boldsymbol{F}\boldsymbol{F}^\top}\omega_1 = S\omega, &\quad \omega'_2 \equiv S'_{\boldsymbol{F}\boldsymbol{F}^\top}\omega'_1 = S'\omega',\\
\mathrm{df}_2 \equiv \mathrm{tr}[\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+&i\omega_2)^{-1}(\boldsymbol{\Sigma}+i\omega'_2)^{-1}].
\end{aligned}
\tag{16}
$$

Then, the weak deterministic equivalents for df_1 and df'_1 can be written as

$$
\begin{aligned}
\mathrm{df}_1 \equiv \mathrm{df}^1_{\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}}(\omega) \simeq \mathrm{df}^1_{\boldsymbol{F}\boldsymbol{F}^\top\boldsymbol{\Sigma}}(\omega_1) \simeq \mathrm{df}^1_{\boldsymbol{\Sigma}}(\omega_2),\\
\mathrm{df}'_1 \equiv \mathrm{df}^1_{\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}}(\omega') \simeq \mathrm{df}^1_{\boldsymbol{F}\boldsymbol{F}^\top\boldsymbol{\Sigma}}(\omega'_1) \simeq \mathrm{df}^1_{\boldsymbol{\Sigma}}(\omega'_2),
\end{aligned}
\tag{17}
$$

where we average first over the randomness in $\hat{\boldsymbol{\Sigma}}$ and then $\boldsymbol{F}\boldsymbol{F}^\top$. With this notation fixed, we have the risk asymptotics

$$
\begin{aligned}
\hat{R}(t) &\simeq \int_\omega e^{2i\omega t}\frac{S_{\boldsymbol{F}\boldsymbol{F}^\top}\bar{\boldsymbol{w}}^\top\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+i\omega_2)^{-1}\bar{\boldsymbol{w}}}{1-\chi\mathrm{Tr}[\boldsymbol{\Sigma}]+i\omega\chi D\mathrm{df}_1},\\
R(t) &\simeq \int_{\omega,\omega'} e^{i(\omega+\omega')t}\mathcal{F}(\omega,\omega')\\
&\quad + \int_\omega e^{2i\omega t}\frac{\chi\mathrm{Tr}[\boldsymbol{\Sigma}]-i\omega_1\chi D\mathrm{df}_1}{1-\chi\mathrm{Tr}[\boldsymbol{\Sigma}]+i\omega\chi D\mathrm{df}_1}S_{\boldsymbol{F}\boldsymbol{F}^\top}\bar{\boldsymbol{w}}^\top\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+i\omega_2)^{-1}\bar{\boldsymbol{w}}.
\end{aligned}
\tag{18}
$$

Here, the forcing term $\mathcal{F}(\omega,\omega')$ is given by

$$
\begin{aligned}
\mathcal{F}(i\omega,i\omega') \simeq \frac{SS'}{1-\gamma_1}\Bigg[&\bar{\boldsymbol{w}}^\top(i\omega_2+\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}(i\omega'_2+\boldsymbol{\Sigma})^{-1}\bar{\boldsymbol{w}}\\
&+\bar{\boldsymbol{w}}^\top(i\omega_2+\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}^2(i\omega'_2+\boldsymbol{\Sigma})^{-1}\bar{\boldsymbol{w}}\\
&\times\frac{\frac{D}{N}\,\mathrm{tr}[(i\omega_2+\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}(i\omega'_2+\boldsymbol{\Sigma})^{-1}]}{1-\frac{D}{N}\mathrm{df}_2(i\omega_2,i\omega'_2)}\Bigg].
\end{aligned}
\tag{19}
$$

and

$$
\gamma_1 \simeq \frac{D}{P}\mathrm{df}_2 + \frac{D}{P}\frac{D}{N}(i\omega_2)(i\omega'_2)\frac{\mathrm{tr}[(i\omega_2+\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}(i\omega'_2+\boldsymbol{\Sigma})^{-1}]^2}{1-\frac{D}{N}\mathrm{df}_2}.
\tag{20}
$$

This is the main result of our work, which recovers and extends earlier results in [8, 24]. Using similar techniques, we show that the SGD kernel terms $\mathcal{K}$ and $\hat{\mathcal{K}}$ have the following single variable Fourier transforms

$$
\mathcal{K}(\omega) = \chi\mathrm{Tr}\boldsymbol{\Sigma}^2\left(\boldsymbol{\Sigma}+i\omega_2\right)^{-1} + \chi\frac{i\omega_2}{N}\mathrm{Tr}\left[\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}+i\omega_2\right)^{-1}\right]^2
\tag{21}
$$

$$
\hat{\mathcal{K}}(\omega) = \mathcal{K}(\omega) + \chi\frac{i\omega_1}{P}\mathrm{Tr}\left[\boldsymbol{\Sigma}\left(\boldsymbol{\Sigma}+i\omega_2\right)^{-1}\right]^2,
\tag{22}
$$

as we derive in Appendix F. Combining our result for the forcing term $\mathcal{F}$ and SGD kernels $\mathcal{K},\hat{\mathcal{K}}$, we obtain our above formulas for the train and test risks $\hat{R}(t), R(t)$.

5. Conclusion

We have derived a class of two-point deterministic equivalents of random matrices. Using this, we have been able to provide sharp asymptotics for the training and generalization performance of a variety of linear models. Our results include both statics and dynamics. In all settings, we see that the S-transform of free probability plays a key role. Several of the results have been obtained in prior literature using either one-point equivalents in random matrix theory [24] or via dynamical mean field theory [8]. Our approach provides a novel diagrammatic derivation of this two-frequency resolvent correlation that is key to capturing the non-commutative dynamics which arises in random feature models.

Acknowledgements

The authors are grateful to Bruno Loureiro, Hamza Chaudhry, Alex Wei, and Jamie Sully for discussions on random matrix theory. We also thank Benjamin Ruben for helpful comments on a previous version of this manuscript.

AA and C. Pehlevan were supported by NSF Award DMS-2134157 and NSF CAREER Award IIS-2239780. B.B. is supported by a Google PhD Fellowship. JAZV is supported by the Office of the Director of the National Institutes of Health under Award Number DP5OD037354. The content is solely the responsibility of the authors and does not necessarily represent the official views of the National Institutes of Health. JAZV is further supported by a Junior Fellowship from the Harvard Society of Fellows. C. Pehlevan is further supported by a Sloan Research Fellowship. C. Paquette is a Canadian Institute for Advanced Research (CIFAR) AI chair, Quebec AI Institute (MILA) and a Sloan Research Fellow in Computer Science (2024). C. Paquette was supported by a Discovery Grant from the Natural Science and Engineering Research Council (NSERC) of Canada, NSERC CREATE grant Interdisciplinary Math and Artificial Intelligence Program (INTER-MATH-AI), Google research grant, and Fonds de recherche du Québec - Nature et technologies (FRQNT) New University Researcher's Start-Up Program. This research is based on work supported by the CIFAR Pan-Canadian AI Strategy through a Catalyst award. Additional revenues related to this work: C. Paquette has 20% part-time employment at Google DeepMind. This work has been made possible in part by a gift from the Chan Zuckerberg Initiative Foundation to establish the Kempner Institute for the Study of Natural and Artificial Intelligence.

References

[1] Ben Adlam and Jeffrey Pennington. The neural tangent kernel in high dimensions: Triple descent and a multi-scale theory of generalization. In *International Conference on Machine Learning*, pages 74–84. PMLR, 2020.

[2] Ibrahim M Alabdulmohsin, Xiaohua Zhai, Alexander Kolesnikov, and Lucas Beyer. Getting ViT in shape: Scaling laws for compute-optimal model design. *Advances in Neural Information Processing Systems*, 36, 2024.

[3] Alexander Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Risk and cross validation in ridge regression with correlated samples. *arXiv preprint arXiv:2408.04607*, 2024.

[4] Alexander Atanasov, Jacob A Zavatone-Veth, and Cengiz Pehlevan. Scaling and renormalization in high-dimensional regression. *arXiv preprint arXiv:2405.00592*, 2024.

[5] Francis Bach. High-dimensional analysis of double descent for linear regression with random projections. *SIAM Journal on Mathematics of Data Science*, 6(1):26–50, 2024.

[6] Gregor Bachmann, Sotiris Anagnostidis, and Thomas Hofmann. Scaling MLPs: A tale of inductive bias. *Advances in Neural Information Processing Systems*, 36, 2024.

[7] Blake Bordelon and Cengiz Pehlevan. Learning curves for SGD on structured features. In *International Conference on Learning Representations*, 2022. URL https://openreview.net/forum?id=WPI2vbkAl3Q.

[8] Blake Bordelon, Alexander Atanasov, and Cengiz Pehlevan. A dynamical model of neural scaling laws. In *Forty-first International Conference on Machine Learning*, 2024.

[9] Abdulkadir Canatar, Blake Bordelon, and Cengiz Pehlevan. Out-of-distribution generalization in kernel regression. In M. Ranzato, A. Beygelzimer, Y. Dauphin, P.S. Liang, and J. Wortman Vaughan, editors, *Advances in Neural Information Processing Systems*, volume 34, pages 12600–12612. Curran Associates, Inc., 2021. URL https://proceedings.neurips.cc/paper_files/paper/2021/file/691dcb1d65f31967a874d18383b9da75-Paper.pdf.

[10] Katie Everett, Lechao Xiao, Mitchell Wortsman, Alexander A Alemi, Roman Novak, Peter J Liu, Izzeddin Gur, Jascha Sohl-Dickstein, Leslie Pack Kaelbling, Jaehoon Lee, et al. Scaling exponents across parameterizations and optimizers. *arXiv preprint arXiv:2407.05872*, 2024.

[11] Behrooz Ghorbani, Orhan Firat, Markus Freitag, Ankur Bapna, Maxim Krikun, Xavier Garcia, Ciprian Chelba, and Colin Cherry. Scaling laws for neural machine translation. *arXiv preprint arXiv:2109.07740*, 2021.

[12] Mitchell A Gordon, Kevin Duh, and Jared Kaplan. Data and parameter scaling laws for neural machine translation. In *Proceedings of the 2021 Conference on Empirical Methods in Natural Language Processing*, pages 5915–5922, 2021.

[13] Danny Hernandez, Jared Kaplan, Tom Henighan, and Sam McCandlish. Scaling laws for transfer. *arXiv preprint arXiv:2102.01293*, 2021.

[14] Danny Hernandez, Tom Brown, Tom Conerly, Nova DasSarma, Dawn Drain, Sheer El-Showk, Nelson Elhage, Zac Hatfield-Dodds, Tom Henighan, Tristan Hume, et al. Scaling laws and interpretability of learning from repeated data. *arXiv preprint arXiv:2205.10487*, 2022.

[15] Joel Hestness, Sharan Narang, Newsha Ardalani, Gregory Diamos, Heewoo Jun, Hassan Kianinejad, Md Mostofa Ali Patwary, Yang Yang, and Yanqi Zhou. Deep learning scaling is predictable, empirically. *arXiv preprint arXiv:1712.00409*, 2017.

[16] Jordan Hoffmann, Sebastian Borgeaud, Arthur Mensch, Elena Buchatskaya, Trevor Cai, Eliza Rutherford, Diego de Las Casas, Lisa Anne Hendricks, Johannes Welbl, Aidan Clark, et al. Training compute-optimal large language models. *arXiv preprint arXiv:2203.15556*, 2022.

[17] Jared Kaplan, Sam McCandlish, Tom Henighan, Tom B Brown, Benjamin Chess, Rewon Child, Scott Gray, Alec Radford, Jeffrey Wu, and Dario Amodei. Scaling laws for neural language models. *arXiv preprint arXiv:2001.08361*, 2020.

[18] Licong Lin, Jingfeng Wu, Sham M Kakade, Peter L Bartlett, and Jason D Lee. Scaling laws in linear regression: Compute, parameters, and data. *arXiv preprint arXiv:2406.08466*, 2024.

[19] Bruno Loureiro, Cedric Gerbelot, Hugo Cui, Sebastian Goldt, Florent Krzakala, Marc Mezard, and Lenka Zdeborová. Learning curves of generic features maps for realistic datasets with a teacher-student model. *Advances in Neural Information Processing Systems*, 34:18137–18151, 2021.

[20] Niklas Muennighoff, Alexander Rush, Boaz Barak, Teven Le Scao, Nouamane Tazi, Aleksandra Piktus, Sampo Pyysalo, Thomas Wolf, and Colin A Raffel. Scaling data-constrained language models. *Advances in Neural Information Processing Systems*, 36, 2024.

[21] Courtney Paquette, Kiwon Lee, Fabian Pedregosa, and Elliot Paquette. Sgd in the large: Average-case analysis, asymptotics, and stepsize criticality. In Mikhail Belkin and Samory Kpotufe, editors, *Proceedings of Thirty Fourth Conference on Learning Theory*, volume 134 of *Proceedings of Machine Learning Research*, pages 3548–3626. PMLR, 08 2021. URL https://proceedings.mlr.press/v134/paquette21a.html.

[22] Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Implicit regularization or implicit conditioning? exact risk trajectories of sgd in high dimensions. In S. Koyejo, S. Mohamed, A. Agarwal, D. Belgrave, K. Cho, and A. Oh, editors, *Advances in Neural Information Processing Systems*, volume 35, pages 35984–35999. Curran Associates, Inc., 2022. URL https://proceedings.neurips.cc/paper_files/paper/2022/file/e9d89428e0ef0a70913845b3ae812ee0-Paper-Conference.pdf.

[23] Courtney Paquette, Elliot Paquette, Ben Adlam, and Jeffrey Pennington. Implicit regularization or implicit conditioning? exact risk trajectories of sgd in high dimensions. *Advances in Neural Information Processing Systems*, 35:35984–35999, 2022.

[24] Elliot Paquette, Courtney Paquette, Lechao Xiao, and Jeffrey Pennington. 4 + 3 phases of compute-optimal neural scaling laws. *arXiv preprint arXiv:2405.15074*, 2024.

[25] Pratik Patil, Jin-Hong Du, and Ryan J. Tibshirani. Optimal ridge regularization for out-of-distribution prediction. *arXiv*, 2024. URL https://arxiv.org/abs/2404.01233.

[26] Marc Potters and Jean-Philippe Bouchaud. *A first course in random matrix theory: for physicists, engineers and data scientists*. Cambridge University Press, 2020.

[27] Andrew Strominger. *Lectures on the infrared structure of gravity and gauge theory*. Princeton University Press, 2018.

[28] Maksim Velikanov, Denis Kuznedelev, and Dmitry Yarotsky. A view of mini-batch sgd via generating functions: conditions of convergence, phase transitions, benefit from negative momenta. *arXiv preprint arXiv:2206.11124*, 2022.

[29] Jacob A Zavatone-Veth and Cengiz Pehlevan. Learning curves for deep structured Gaussian feature models. In *Advances in Neural Information Processing Systems*, 2023.

[30] Xiaohua Zhai, Alexander Kolesnikov, Neil Houlsby, and Lucas Beyer. Scaling vision transformers. In *Proceedings of the IEEE/CVF conference on computer vision and pattern recognition*, pages 12104–12113, 2022.

Appendix A. A Reduced Model for SGD Risk Dynamics

We now derive a reduced model for the dynamics of the training and generalization error under SGD, following previous works that argue this reduced model should be asymptotically equivalent in high dimensions [7, 21–24]. We will not give a complete proof of this equivalence here, but instead reference conditions under which prior art gives either a rigorous proof or compelling evidence that it holds.

We first observe that all quantities of interest—namely the population loss R_t, the training loss $\hat{R}_t$, and the loss on a given batch $\hat{R}_{\mathcal{B}_t}$—can be written as quadratic forms in the weight discrepancy $\Delta \boldsymbol{w}_t \equiv \bar{\boldsymbol{w}} - \boldsymbol{w}_t$:

$$R_t = \Delta \boldsymbol{w}_t^\top \boldsymbol{\Sigma} \Delta \boldsymbol{w}_t, \quad \hat{R}_t = \Delta \boldsymbol{w}_t^\top \hat{\boldsymbol{\Sigma}} \Delta \boldsymbol{w}_t, \quad \hat{R}_{\mathcal{B}_t} = \Delta \boldsymbol{w}_t^\top \hat{\boldsymbol{\Sigma}}_t \Delta \boldsymbol{w}_t. \tag{23}$$

From the equations for SGD we have

$$\begin{aligned} \boldsymbol{v}_{t+1} &= \boldsymbol{v}_t - \eta \nabla_{\boldsymbol{v}} \hat{R}_{\mathcal{B}_t} = v_t + \eta \boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}}_t \Delta \boldsymbol{w}_t \\ \Rightarrow \Delta \boldsymbol{w}_{t+1} &= \Delta \boldsymbol{w}_t - \eta \boldsymbol{F}\boldsymbol{F}^\top \underbrace{\hat{\boldsymbol{\Sigma}}_t \Delta \boldsymbol{w}_t}_{\boldsymbol{g}_t}. \end{aligned} \tag{24}$$

To derive a recursion for the second moment of the weight discrepancy $\Delta \boldsymbol{w}_t$, we will make use of a Gaussian approximation of the features $\boldsymbol{x}_{\mu,t}$. In [7, 28], this approximation was shown to give a reasonable set of learning curves even if the true features are non-Gaussian and dimension free. Under this approximation, we can compute the first two moments of $\boldsymbol{g}_t$:

$$\mathbb{E}_{\mathcal{B}_t} \boldsymbol{g}_t = \hat{\boldsymbol{\Sigma}} \Delta \boldsymbol{w}_t, \tag{25}$$

$$\mathbb{E}_{\mathcal{B}_t} \boldsymbol{g}_t \boldsymbol{g}_t^\top = \hat{\boldsymbol{\Sigma}} \Delta \boldsymbol{w}_t \Delta \boldsymbol{w}_t^\top \hat{\boldsymbol{\Sigma}} + \frac{1}{B} \hat{\boldsymbol{\Sigma}} \Delta \boldsymbol{w}_t \Delta \boldsymbol{w}_t^\top \hat{\boldsymbol{\Sigma}} + \frac{1}{B} \hat{\boldsymbol{\Sigma}} \underbrace{\Delta \boldsymbol{w}_t^\top \hat{\boldsymbol{\Sigma}} \Delta \boldsymbol{w}_t}_{\hat{R}_t}. \tag{26}$$

Because R, $\hat{R}$ are quadratic functions of $\Delta \boldsymbol{w}_t$, then following [7], it is sufficient to track $\Delta \boldsymbol{w}_t \Delta \boldsymbol{w}_t^\top$:

$$\begin{aligned} \Delta \boldsymbol{w}_{t+1} \Delta \boldsymbol{w}_{t+1}^\top = \Delta \boldsymbol{w}_t \Delta \boldsymbol{w}_t^\top &- \eta \Delta \boldsymbol{w}_t \boldsymbol{g}_t \boldsymbol{F}\boldsymbol{F}^\top \\ &- \eta \boldsymbol{F}\boldsymbol{F}^\top \boldsymbol{g}_t \Delta \boldsymbol{w}_t^\top + \eta^2 \boldsymbol{F}\boldsymbol{F}^\top \boldsymbol{g}_t \boldsymbol{g}_t^\top \boldsymbol{F}\boldsymbol{F}^\top. \end{aligned} \tag{27}$$

Taking expectations over the batch and writing $\boldsymbol{C}_t \equiv \mathbb{E}_{\mathcal{B}_t} \Delta \boldsymbol{w}_t \Delta \boldsymbol{w}_t^\top$ and $\chi = \eta/B$ for the **SGD temperature** yields:

$$\begin{aligned} \boldsymbol{C}_{t+1} = (1 - \eta \boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}}) \boldsymbol{C}_t (1 - \eta \hat{\boldsymbol{\Sigma}} \boldsymbol{F}\boldsymbol{F}^\top) \\ + \eta\chi \boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{C}_t \hat{\boldsymbol{\Sigma}} \boldsymbol{F}\boldsymbol{F}^\top + \eta\chi \boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{F}\boldsymbol{F}^\top \mathrm{Tr}[\boldsymbol{C}_t \hat{\boldsymbol{\Sigma}}] \end{aligned} \tag{28}$$

In what follows, we will drop the middle term in Equation (28). Depending on the structure of the data, this term either (1) explicitly vanishes in the high-dimensional $D \to \infty$ limit, or (2) contributes negligibly after sufficient time t. In the first case where $D \to \infty$ with $\frac{1}{D}\mathrm{Tr}(\boldsymbol{\Sigma}) = \Theta(1)$, maintaining stable dynamics with $\eta = \Theta(1)$ requires choosing large enough batch size so that $\chi = \Theta(D^{-1})$. With this choice, the two terms generated by SGD effects scale as

$$\eta\chi \boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{C}_t \hat{\boldsymbol{\Sigma}} \boldsymbol{F}\boldsymbol{F}^\top = \Theta(D^{-1}) \ , \ \eta\chi \boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{F}\boldsymbol{F}^\top \mathrm{Tr}[\boldsymbol{C}_t \hat{\boldsymbol{\Sigma}}] = \Theta(1), \tag{29}$$

justifying neglecting the left term. Alternatively, if the features are dimension free so that $\mathrm{Tr}\boldsymbol{\Sigma} < \infty$ as $D \to \infty$, then we note that the spectrum of $\boldsymbol{\Sigma}$ must decay sufficiently rapidly. The projection of these terms along the k-th population eigendirection $\boldsymbol{v}_k$ are

$$\boldsymbol{v}_k^\top \boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{C}_t \hat{\boldsymbol{\Sigma}} \boldsymbol{F}\boldsymbol{F}^\top \boldsymbol{v}_k \approx \lambda_k^2 \ , \ \boldsymbol{v}_k^\top \boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}} \boldsymbol{F}\boldsymbol{F}^\top \boldsymbol{v}_k \mathrm{Tr}[\boldsymbol{C}_t \hat{\boldsymbol{\Sigma}}] \approx \lambda_k \hat{\mathcal{L}}(t) \tag{30}$$

where $\hat{\mathcal{L}}(t) = \mathrm{Tr}[\boldsymbol{C}_t \hat{\boldsymbol{\Sigma}}]$ is the training loss. For a small eigenvalue $\lambda_k \ll 1$ the first term will be dominated by the second term since it is quadratic in the small eigenvalue. Based on the two cases above, we thus consider the simplified dynamics

$$\boldsymbol{C}_{t+1} = (1 - \eta \boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}})\boldsymbol{C}_t(1 - \eta \hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top) + \eta\chi \boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top \mathrm{Tr}[\boldsymbol{C}_t\hat{\boldsymbol{\Sigma}}]. \tag{31}$$

We can now take the $\eta \to 0$ limit of this equation while keeping χ constant as in [7]. This will then include the SGD noise contributions to the training dynamics. By utilizing an integrating factor, we obtain the following differential equation:

$$\frac{d}{dt}\left[e^{\eta t \boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}}}\boldsymbol{C}_t e^{\eta t \hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top}\right] = \chi\, e^{2t\boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}}}\boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top \mathrm{Tr}[\boldsymbol{C}_t\hat{\boldsymbol{\Sigma}}]. \tag{32}$$

On the right-hand side we have applied the push-through identity to the matrix exponential. Integrating this, and using that $\boldsymbol{C}_0 = \bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^\top$, we obtain the Volterra equation of [21, 22, 24]:

$$\boldsymbol{C}_t \simeq e^{-\eta \boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}}}\bar{\boldsymbol{w}}\bar{\boldsymbol{w}}^\top e^{-\eta \hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top} + \chi \int_0^t e^{-2(t-s)\boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}}}\boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top \mathrm{Tr}[\boldsymbol{C}_s\hat{\boldsymbol{\Sigma}}]ds \tag{33}$$

Tracing against $\hat{\boldsymbol{\Sigma}}$, $\boldsymbol{\Sigma}$ gives the evolution for the training and test losses respectively:

$$\boxed{\hat{R}_t = \underbrace{\bar{\boldsymbol{w}}^\top e^{-2t\hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top}\hat{\boldsymbol{\Sigma}}\bar{\boldsymbol{w}}}_{\hat{\mathcal{F}}(t)} + \chi \int_0^t \underbrace{\mathrm{Tr}[e^{-2(t-s)\boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}}}(\boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}})^2]}_{\hat{\mathcal{K}}(t-s)}\hat{R}_s ds,} \tag{34}$$

$$\boxed{\begin{aligned} R_t = &\underbrace{\bar{\boldsymbol{w}}^\top e^{-t\hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top}\boldsymbol{\Sigma} e^{-t\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}}\bar{\boldsymbol{w}}}_{\mathcal{F}(t)} \\ &+ \chi \int_0^t \underbrace{\mathrm{Tr}[e^{-2(t-s)\boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}}}\boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top \boldsymbol{\Sigma}]}_{\mathcal{K}(t-s)}\hat{R}_s ds. \end{aligned}} \tag{35}$$

In both equations, the first term on the right hand side is referred to as the **forcing term**, while the second term we will refer to as the response or **kernel** term. It is the second term that is due to SGD noise, and would go away in the limit of $\chi \to 0$. The kernel term consists of a convolution of the population risk with the train and test kernels $\hat{\mathcal{K}}, \mathcal{K}$ respectively. This notation is adopted from [24].

A.1. Forcing Terms

The forcing term in the equation for the generalization error will require a novel two-point deterministic equivalent to be derived. We first consider the more general quantity:

$$\mathcal{F}(t,t') = \Delta \boldsymbol{w}(t)^\top \boldsymbol{\Sigma} \Delta \boldsymbol{w}(t') = \bar{\boldsymbol{w}}^\top e^{-t\hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top}\boldsymbol{\Sigma} e^{-t'\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}}\bar{\boldsymbol{w}}. \tag{36}$$

We want the diagonal $t = t'$ of this function. We first Fourier transform in t, t' separately to obtain

$$\begin{aligned} \mathcal{F}(t) &= \int_{\omega,\omega'} e^{it(\omega+\omega')}\mathcal{F}(\omega,\omega') \\ \mathcal{F}(\omega,\omega') &= \bar{\boldsymbol{w}}^\top(\hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top + i\omega)^{-1}\boldsymbol{\Sigma}(\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}} + i\omega')^{-1}\bar{\boldsymbol{w}}. \end{aligned} \tag{37}$$

where we adopt the shorthand

$$\int_\omega (\cdot) = \frac{1}{2\pi}\int (\cdot)\, d\omega \quad \text{and} \quad \int_{\omega,\omega'} (\cdot) = \frac{1}{(2\pi)^2}\int (\cdot)\, d\omega\, d\omega' \tag{38}$$

for integrals over Fourier space. We see that in Fourier space, all of the randomness enters through the product of two resolvents evaluated at different "imaginary ridges", $i\omega, i\omega'$. The main technical goal of this note is to provide sharp asymptotics for this product, which we do in Section D.

A.2. Kernel Terms

Denoting temporal convolution by $\hat{\mathcal{K}} \star \hat{R}$, we can rewrite Equation (6) as

$$\begin{aligned}\hat{R} &= \hat{\mathcal{F}} + \chi \hat{\mathcal{K}} \star \hat{R} \\ &= \hat{\mathcal{F}} + \chi \hat{\mathcal{K}} \star \hat{\mathcal{F}} + \chi^2 \hat{\mathcal{K}} \star \hat{\mathcal{K}} \star \hat{\mathcal{F}} + \dots \\ &= (1 - \chi \hat{\mathcal{K}})^{-1} \star \hat{\mathcal{F}}.\end{aligned} \tag{39}$$

Here, the last two lines are to be understood formally. For the test risk R given in (7), an identical formal equation holds with $\hat{\mathcal{F}}$ replaced by $\mathcal{F}$. In order for SGD to be stable, we need a constraint that $\chi \|\mathcal{K}\| < 1$ in operator norm, such that this Neumann series converges.

Appendix B. Recovering statics

The long time limit of both the forcing function and the kernel function can be studied in Fourier space a well. In the single frequency setting, we require that all poles of the function in question lie in *either* the upper half-plane or at $\omega = 0$ on the real line. We see that the residues of any pole will be multiplied by the factor $e^{i\omega t}$. This goes to zero as $t \to \infty$ for any ω with $\mathrm{Im}\omega > 0$. The remaining poles are on the real line. Poles away from zero would lead to oscillatory behavior at infinite time, and do not appear in the quantities that we treat. The same argument applies jointly to ω, ω'.

Assuming the remaining joint pole at $\omega, \omega' = 0$ is simple in both variables, we get that

$$\lim_{t \to \infty} \mathcal{F}(t) = \lim_{\omega, \omega' \to 0} (i\omega)(i\omega')\mathcal{F}(i\omega, i\omega'). \tag{40}$$

This is known in the Laplace transform literature as the **final value theorem**, or more generally in physics as a **soft limit**. The role of such soft limits in recovering static effects, also known as DC effects or "memory effects", has been highlighted in recent physics literature [27].

Appendix C. Random matrix notation

C.1. Degrees of Freedom

In what follows, we will use tr to denote the normalized trace. For an $N \times N$ matrix $\boldsymbol{A}$, this is

$$\mathrm{tr}[\boldsymbol{A}] = \frac{1}{N}\mathrm{Tr}[\boldsymbol{A}]. \tag{41}$$

The key quantities that emerge in the study of ridge regression in the case of statics are the *degrees of freedom*. We define $\mathrm{df}^1_{\boldsymbol{A}}$ and $\mathrm{df}^2_{\boldsymbol{A}}$ as follows

$$\mathrm{df}^1_{\boldsymbol{A}}(\lambda) \equiv \mathrm{tr}[\boldsymbol{A}(\boldsymbol{A}+\lambda)^{-1}], \quad \mathrm{df}^2_{\boldsymbol{A}}(\lambda) \equiv \mathrm{tr}[\boldsymbol{A}^2(\boldsymbol{A}+\lambda)^{-2}]. \tag{42}$$

We will also write:

$$\mathrm{df}^2_{\boldsymbol{A}}(\lambda, \lambda') \equiv \mathrm{tr}[\boldsymbol{A}^2(\boldsymbol{A}+\lambda)^{-1}(\boldsymbol{A}+\lambda')^{-1}], \tag{43}$$

$$\mathrm{df}^2_{\boldsymbol{A},\boldsymbol{A}'}(\lambda, \lambda') \equiv \mathrm{tr}[\boldsymbol{A}\boldsymbol{A}'(\boldsymbol{A}+\lambda)^{-1}(\boldsymbol{A}+\lambda')^{-1}]. \tag{44}$$

Notice the last definition is *not* symmetric in $\boldsymbol{A}, \boldsymbol{A}'$. We also define the *teacher-weighted* degrees of freedom as:

$$\mathrm{tf}^1_{\boldsymbol{A},\bar{\boldsymbol{w}}}(\lambda) \equiv \bar{\boldsymbol{w}}^\top \boldsymbol{A}(\boldsymbol{A}+\lambda)^{-1}\bar{\boldsymbol{w}}, \quad \mathrm{tf}^2_{\boldsymbol{A},\bar{\boldsymbol{w}}}(\lambda) \equiv \bar{\boldsymbol{w}}^\top \boldsymbol{A}^2(\boldsymbol{A}+\lambda)^{-2}\bar{\boldsymbol{w}}. \tag{45}$$

These are defined as in [3, 4, 29].

C.2. S-transform

Throughout this work, we will find that many key quantities can be expressed in terms of the S-transform of free probability. For a random matrix $\boldsymbol{A}$ drawn from some ensemble, we define the S-transform of the ensemble to be a function of the formal variable df. Letting $\mathrm{df}_{\boldsymbol{A}}^{(-1)}(\mathrm{df})$ be the functional inverse of $\mathrm{df}_{\boldsymbol{A}}^1$, namely $\mathrm{df}_{\boldsymbol{A}}^{(-1)}(\mathrm{df}_{\boldsymbol{A}}^1(\lambda)) = \lambda$, we have:

$$S_{\boldsymbol{A}}(\mathrm{df}) \equiv \frac{1-\mathrm{df}}{\mathrm{df}\,\mathrm{df}_{\boldsymbol{A}}^{-1}(\mathrm{df})}. \tag{46}$$

A consequence of this is that for all λ:

$$\mathrm{df}_{\boldsymbol{A}}^1(\lambda) = \frac{1}{1+S_{\boldsymbol{A}}(\mathrm{df}_{\boldsymbol{A}}^1(\lambda))\lambda}. \tag{47}$$

The S-transform has the property that when two matrices $\boldsymbol{A}$, $\boldsymbol{B}$ are **free** of one another, one has the following property

$$S_{\boldsymbol{A}*\boldsymbol{B}}(\mathrm{df}) = S_{\boldsymbol{A}}(\mathrm{df})S_{\boldsymbol{B}}(\mathrm{df}). \tag{48}$$

Here $\boldsymbol{A}*\boldsymbol{B} \equiv \boldsymbol{A}^{1/2}\boldsymbol{B}\boldsymbol{A}^{1/2}$ for $\boldsymbol{A}^{1/2}$ the principal matrix square root of $\boldsymbol{A}$. This is sometimes called the free product. A consequence of (47) and (48) is that for $\boldsymbol{A}, \boldsymbol{B}$ free of one another, one has:

$$\mathrm{df}_{\boldsymbol{A}*\boldsymbol{B}}^1(\lambda) = \mathrm{df}_{\boldsymbol{A}}^1(\kappa) \tag{49}$$

Here, κ is known as the **resolution** or **signal capture threshold**, and can be calculated in two different ways. It can be calculated **empirically** as

$$\kappa = \lambda S_{\boldsymbol{B}}(\mathrm{df}_{\boldsymbol{A}*\boldsymbol{B}}(\lambda)), \tag{50}$$

or **omnisciently** via the self-consistent equation:

$$\kappa = \lambda S_{\boldsymbol{B}}(\mathrm{df}_{\boldsymbol{A}}(\kappa)). \tag{51}$$

Equation (49) is known as a **subordination relation** or as a **weak deterministic equivalence**. Equations (50) and (51) are equivalent precisely because equation (49) holds.

The S-transform is particularly useful as it arises also in **strong deterministic equivalence**, where one can write

$$\boldsymbol{A}\boldsymbol{B}(\lambda+\boldsymbol{A}\boldsymbol{B})^{-1} \simeq \boldsymbol{A}(\boldsymbol{A}+\kappa)^{-1}. \tag{52}$$

This is the non-traced form of the weak deterministic equivalence of equation (49). Here, by $\boldsymbol{M}_1 \simeq \boldsymbol{M}_2$ we mean that $\mathrm{tr}[\boldsymbol{M}_1\boldsymbol{\Theta}]/\,\mathrm{tr}[\boldsymbol{M}_2\boldsymbol{\Theta}] \to 1$ as $N \to \infty$ for any sequence of test matrices $\boldsymbol{\Theta}$ of finite spectral norm [4, 5, 26].

Appendix D. Two-Point Deterministic Equivalence

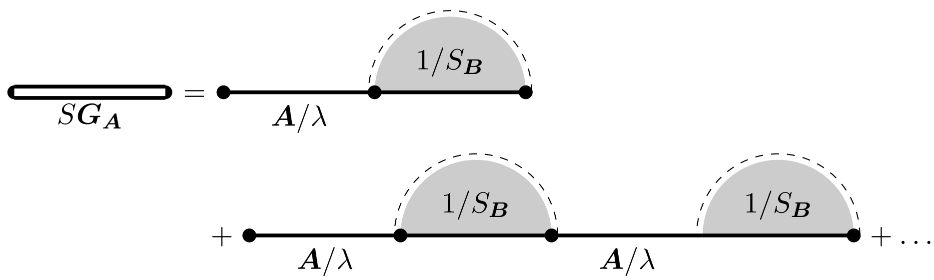
Let $\boldsymbol{A}, \boldsymbol{M}$ be deterministic and $\boldsymbol{B}$ be an isotropic multiplicative noise source. All are $N \times N$ matrices and $\boldsymbol{B}$ is free of $\boldsymbol{A}$, $\boldsymbol{M}$. We will eventually specialize the case in which $\boldsymbol{B}$ is a white Wishart matrix, but our derivations hold for general $\boldsymbol{B}$ satisfying the freedom assumption. We are interested in finding a deterministic equivalent for the following expression:

$$(\lambda+\boldsymbol{A}\boldsymbol{B})^{-1}\boldsymbol{M}(\lambda'+\boldsymbol{B}\boldsymbol{A})^{-1}. \tag{53}$$

We call this a **two point resolvent**, by analogy to similar quantities in field theory that involve the insertion of an operator, in this case $(\lambda+\boldsymbol{A}\boldsymbol{B})^{-1}$, at two different points, in this case λ, λ'. We will use the following shorthand to simplify our final equations:

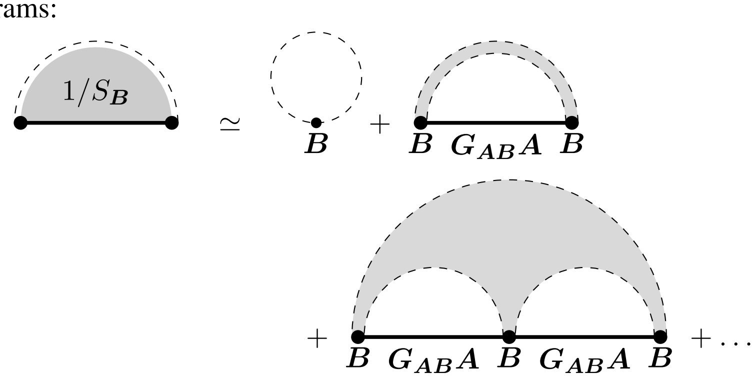
$$\begin{aligned} S_{\boldsymbol{B}} &= S_{\boldsymbol{B}}(-\mathrm{df}^1_{\boldsymbol{AB}}(\lambda)), & S'_{\boldsymbol{B}} &= S_{\boldsymbol{B}}(-\mathrm{df}^1_{\boldsymbol{AB}}(\lambda')), \\ \kappa &= \lambda S_{\boldsymbol{B}}, & \kappa' &= \lambda' S'_{\boldsymbol{B}}, \\ \boldsymbol{G}_{\boldsymbol{A}} &= (\kappa + \boldsymbol{A})^{-1}, & \boldsymbol{G}'_{\boldsymbol{A}} &= (\kappa' + \boldsymbol{A})^{-1}, \\ g &= \mathrm{tr}[\boldsymbol{G}_{\boldsymbol{A}}], & g' &= \mathrm{tr}[\boldsymbol{G}'_{\boldsymbol{A}}], \\ \boldsymbol{T}_{\boldsymbol{A}} &= \boldsymbol{A}(\kappa + \boldsymbol{A})^{-1}, & \boldsymbol{T}'_{\boldsymbol{A}} &= \boldsymbol{A}(\kappa' + \boldsymbol{A})^{-1}. \end{aligned} \tag{54}$$

We first review a variation of the argument in [4] to evaluate a single $\boldsymbol{G}_{\boldsymbol{A}}$. For more details on the orthogonal averages, the reader is encouraged to first go through Section III of that work. Following that argument, we utilize the freedom of $\boldsymbol{B}$ relative to $\boldsymbol{A}$ to write it as $\boldsymbol{O}\boldsymbol{B}'\boldsymbol{O}^\top$ with $\boldsymbol{B}'$ diagonal, and perform averages over the orthogonal matrix $\boldsymbol{O}$. In performing the average over the orthogonal group, using the language of [4], we can expand the resolvent in terms of a series of *irreducible* diagrams linked together by multiplications by $\boldsymbol{A}/\lambda$. By orthogonal invariance, the irreducible diagrams must be scalars, equal to a value $1/S_{\boldsymbol{B}}$. This re-sums to $S_{\boldsymbol{B}} G_{\boldsymbol{A}}$. Diagrammatically, we write:



(55)

Here, again the language of [4], each irreducible diagram can be expressed as a sum of *fully connected* diagrams:



(56)

$$= \sum_{n=1}^{\infty} \kappa^{(n)}_{\boldsymbol{B}} \mathrm{tr}[\boldsymbol{G}_{\boldsymbol{AB}}\boldsymbol{A}]^{n-1} = \sum_{n=1}^{\infty} \kappa^{(n)}_{\boldsymbol{B}} (S\mathrm{df}_{\boldsymbol{A}})^{n-1}.$$

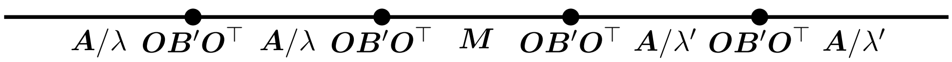
Here, the shaded grey diagrams on the right hand side correspond to averages over the $\boldsymbol{O}$ that are, at leading order, identical to simple Wick contractions, but have subleading "Weingarten" terms that can still contribute to the final results. See section III of [4] for a deeper discussion of this. $\kappa^{(n)}_{\boldsymbol{B}}$ are the **free cumulants** of $\boldsymbol{B}$, as in [4, 26]. There, the R-transform is given by the power series in the formal variable g:

$$R_{\boldsymbol{B}}(g) = \sum_{n=1}^{\infty} \kappa^{(n)}_{\boldsymbol{B}} g^{n-1}. \tag{57}$$

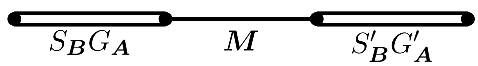
Moreover, the R and S transform are related by the identity that $R_{\boldsymbol{B}}(S_{\boldsymbol{B}}\mathrm{df}) = S_{\boldsymbol{B}}(\mathrm{df})^{-1}$. See [4, 26] for details. We thus have the strong deterministic equivalence from the prior work:

$$G_{\boldsymbol{AB}}(\lambda) \simeq S_{\boldsymbol{B}} G_{\boldsymbol{A}}(\kappa), \quad S_{\boldsymbol{B}} = S_{\boldsymbol{B}}(\mathrm{df}_{\boldsymbol{A}}(\kappa)). \tag{58}$$

Having reviewed this "one-point" deterministic equivalence of prior work, we now extend this approach to evaluate Equation (53). We now expand in $1/\lambda$ and $1/\lambda'$ jointly. Before performing the orthogonal average, the general term will look like:

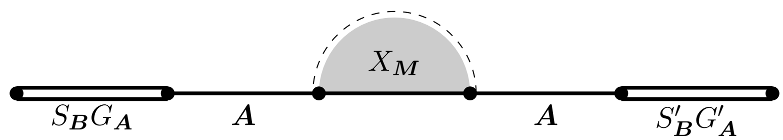


Upon performing the average over $\boldsymbol{O}$, we recall that crossing diagrams do not contribute in the large N limit. From this, we see that there will be two classes of terms in the diagrammatics. The first class is from averages that factorize into expectations over the individual resolvents $\boldsymbol{G}, \boldsymbol{G}'$. That is, the averages over the orthogonal group are performed separately to the left and to the right of $\boldsymbol{M}$. This "disconnected" contribution can be evaluated by appeal to the one-point equivalences separately on the left and right, and thus yields:



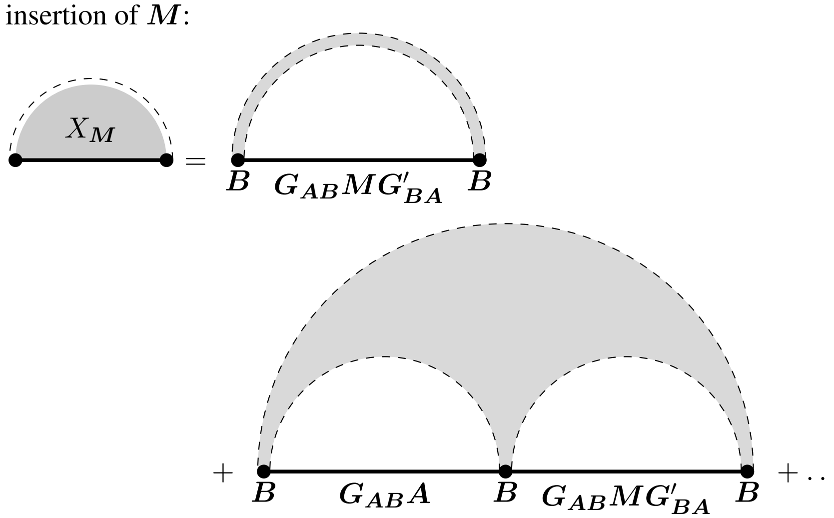
$$= S_{\boldsymbol{B}} S'_{\boldsymbol{B}} \boldsymbol{G_A M G'_A} \tag{59}$$

The second term will involve averages that include a cumulant containing $\boldsymbol{M}$ underneath one of its arcs. Such terms will take the form:



(60)

Just like $S_{\boldsymbol{B}}$, $X_{\boldsymbol{M}}$ is again an expansion of fully connected diagrams, but under one of the arcs is an additional insertion of $\boldsymbol{M}$:



(61)

Importantly, note that the factors of $1/\lambda$, $1/\lambda'$ that do not accompany the $\boldsymbol{A}$ are indeed accounted for. They are absorbed into the $\boldsymbol{G_{AB}}, \boldsymbol{G'_{AB}}$ on both sides of $\boldsymbol{M}$.

In general, one can see that a given term in the expansion of $X_{\boldsymbol{M}}$ is labelled by two positive integers a, b. The first $a-1$ arcs will contain $\boldsymbol{G_{AB} A}$ underneath, followed by a single arc containing $\boldsymbol{G_{AB} M G'_{BA}}$ underneath followed by $b-1$ arcs that will contain $\boldsymbol{A G'_{BA}}$ underneath. This term will have $a+b$ insertions of $\boldsymbol{B}$ and thus will involve the $(a+b)$th cumulant of $\boldsymbol{B}$. *Crucially*, as in the derivation of the S-transform, the nth cumulant only depend on the quantities under each arc *through their traces*. We will denote the product of the first $a-1$ traces by g^{a-1} and the product of the last $b-1$ traces by ${g'}^{b-1}$. One can then write a self-consistent equation for $X_{\boldsymbol{M}}$ by recognizing the quantity under the arc with the $\boldsymbol{M}$ insertion is precisely the trace of the original two-point resolvent we sought to evaluate:

$$\begin{aligned} X_{\boldsymbol{M}} &= S_{\boldsymbol{B}} S'_{\boldsymbol{B}} \sum_{n=1}^{\infty} \sum_{a+b=n} \kappa_{\boldsymbol{B}}^{(n)} g^{a-1} {g'}^{b-1} \operatorname{tr}[G_{\boldsymbol{AB}} \boldsymbol{M} G_{\boldsymbol{AB}}] \\ &= S_{\boldsymbol{B}} S'_{\boldsymbol{B}} R_{\boldsymbol{B}}[g, g'] \Big(\operatorname{tr}[\boldsymbol{G_A M G'_A}] + X_{\boldsymbol{M}} \operatorname{tr}[\boldsymbol{G_A A^2 G'_A}] \Big). \end{aligned} \tag{62}$$

$$R_{\boldsymbol{B}}[g, g'] = \sum_{n=1}^{\infty} \sum_{a+b=n} \kappa_{\boldsymbol{B}}^{(n)} g^{a-1} {g'}^{b-1}. \tag{63}$$

where we have introduced the "mixed" R-transform

$$R_{\boldsymbol{B}}[g, g'] = \sum_{n=1}^{\infty} \sum_{a+b=n} \kappa_{\boldsymbol{B}}^{(n)} g^{a-1} {g'}^{b-1}. \tag{64}$$

We can then solve this self-consistent equation for $X_{\boldsymbol{M}}$:

$$X_{\boldsymbol{M}} = \frac{S_{\boldsymbol{B}} S'_{\boldsymbol{B}} R_{\boldsymbol{B}}[g, g'] \operatorname{tr}[\boldsymbol{G_A M G'_A}]}{1 - S_{\boldsymbol{B}} S'_{\boldsymbol{B}} R_{\boldsymbol{B}}[g, g'] \operatorname{tr}[\boldsymbol{G_A A^2 G'_A}]}. \tag{65}$$

Therefore, we have the general deterministic equivalence

$$(\lambda + \boldsymbol{AB})^{-1} \boldsymbol{M} (\lambda' + \boldsymbol{BA})^{-1} \simeq S_{\boldsymbol{B}} S'_{\boldsymbol{B}} \left[\boldsymbol{G_A M G'_A} + X_M \boldsymbol{G_A A^2 G'_A} \right] \tag{66}$$

with $X_{\boldsymbol{M}}$ as above.

The "mixed R-transform" $R_{\boldsymbol{B}}[g, g']$ simplifies in the case when $\boldsymbol{B}$ is a white Wishart matrix $\boldsymbol{B} = \frac{1}{P} \boldsymbol{X}^\top \boldsymbol{X}$. There, because of the factorization property $\kappa_{\boldsymbol{B}}^{(a+b)} = q \kappa_{\boldsymbol{B}}^{(a)} \kappa_{\boldsymbol{B}}^{(b)}$, one has $R_{\boldsymbol{B}}[g, g'] = q R_{\boldsymbol{B}}[g] R_{\boldsymbol{B}}[g'] = q (S_{\boldsymbol{B}} S'_{\boldsymbol{B}})^{-1}$ with $q = N/P$, so that

$$X_{\boldsymbol{M}} = \frac{q \operatorname{tr}[\boldsymbol{G_A M G'_A}]}{1 - q \mathrm{df}_2(\kappa, \kappa')}, \quad \mathrm{df}_2(\kappa, \kappa') \equiv \operatorname{tr}[\boldsymbol{A^2 G_A G'_A}]. \tag{67}$$

All together we get the final deterministic equivalence in the Wishart case:

$$\boxed{\begin{aligned} &(\lambda + \boldsymbol{AB})^{-1} \boldsymbol{M} (\lambda' + \boldsymbol{BA})^{-1} \\ &\quad \simeq S_{\boldsymbol{B}} S'_{\boldsymbol{B}} \left[\boldsymbol{G_A M G'_A} + \boldsymbol{G_A A^2 G'_A} \frac{q \operatorname{tr}[\boldsymbol{G_A M G'_A}]}{1 - q \mathrm{df}_2(\kappa, \kappa')} \right]. \end{aligned}} \tag{68}$$

The special case of this equivalent for $\lambda = \lambda'$ has appeared in our previous work [3] (there, we worked directly in the Wishart case, for which the diagrammatics directly represent Wick contractions). For Gram Wishart matrices, e.g. $\boldsymbol{B} = \frac{1}{P} \boldsymbol{X} \boldsymbol{X}^\top$ or $\boldsymbol{B} = \frac{1}{D} \boldsymbol{F} \boldsymbol{F}^\top$, the above equations hold but with $q \to 1/q$, since there $\kappa_{a+b} = q^{-1} \kappa_a \kappa_b$. We state a variety of additional variants of the above formula in Appendix D.1.

D.1. All Two-Point Deterministic Equivalences

In this section, we report all variants of the two-point deterministic equivalences. These extend prior equivalences observed in [3, 5] to the case of different ridges λ, λ'. First, we have:

$$\begin{aligned}&(\lambda + \boldsymbol{AB})^{-1}\boldsymbol{M}(\lambda' + \boldsymbol{AB})^{-1}\\ &\quad\simeq S_{\boldsymbol{B}}S'_{\boldsymbol{B}}\left[\boldsymbol{G_A M G'_A} + \boldsymbol{G_A A G'_A}\frac{q\,\mathrm{tr}[\boldsymbol{A G_A M G'_A}]}{1 - q\mathrm{df}_2(\kappa,\kappa')}\right].\end{aligned} \tag{69}$$

One also obtains for $\boldsymbol{A} * \boldsymbol{B} = \boldsymbol{A}^{1/2}\boldsymbol{B}\boldsymbol{A}^{1/2}$ the same deterministic equivalence:

$$\begin{aligned}&(\lambda + \boldsymbol{A} * \boldsymbol{B})^{-1}\boldsymbol{M}(\lambda' + \boldsymbol{A} * \boldsymbol{B})^{-1}\\ &\quad\simeq S_{\boldsymbol{B}}S'_{\boldsymbol{B}}\left[\boldsymbol{G_A M G'_A} + \boldsymbol{G_A A G'_A}\frac{q\,\mathrm{tr}[\boldsymbol{A G_A M G'_A}]}{1 - q\mathrm{df}_2(\kappa,\kappa')}\right].\end{aligned} \tag{70}$$

Additionally, one has:

$$\begin{aligned}&(\lambda + \boldsymbol{BA})^{-1}\boldsymbol{M}(\lambda' + \boldsymbol{AB})^{-1}\\ &\quad\simeq S_{\boldsymbol{B}}S'_{\boldsymbol{B}}\left[\boldsymbol{G_A M G'_A} + \boldsymbol{G_A G'_A}\frac{q\,\mathrm{tr}[\boldsymbol{A}^2\boldsymbol{G_A M G'_A}]}{1 - q\mathrm{df}_2(\kappa,\kappa')}\right].\end{aligned} \tag{71}$$

This was for resolvents. By the definition of $\boldsymbol{T_A}, \boldsymbol{T'_A}$, one immediately gets:

$$\begin{aligned}&\boldsymbol{AB}(\lambda + \boldsymbol{AB})^{-1}\boldsymbol{MAB}(\lambda' + \boldsymbol{AB})^{-1}\\ &\quad\simeq \boldsymbol{T_A M T'_A} + \kappa\kappa'\boldsymbol{G_A A G'_A}\frac{q\,\mathrm{tr}[\boldsymbol{A G_A M G'_A}]}{1 - q\mathrm{df}_2(\kappa,\kappa')}.\end{aligned} \tag{72}$$

One has the same deterministic equivalence upon replacing $\boldsymbol{AB}$ by the free product $\boldsymbol{A} * \boldsymbol{B}$.

$$\begin{aligned}&\boldsymbol{A} * \boldsymbol{B}(\lambda + \boldsymbol{A} * \boldsymbol{B})^{-1}\boldsymbol{MA} * \boldsymbol{B}(\lambda' + \boldsymbol{A} * \boldsymbol{B})^{-1}\\ &\quad\simeq \boldsymbol{T_A M T'_A} + \kappa\kappa'\boldsymbol{G_A A G'_A}\frac{q\,\mathrm{tr}[\boldsymbol{A G_A M G'_A}]}{1 - q\mathrm{df}_2(\kappa,\kappa')}.\end{aligned} \tag{73}$$

Again, for Gram matrices, the above equations hold but with $q \to 1/q$.

By adopting the notation $\hat{\boldsymbol{G}}_{\boldsymbol{A}} = \boldsymbol{G_{AB}}, \hat{\boldsymbol{T}}_{\boldsymbol{A}} = \boldsymbol{T_{AB}}, \gamma = q\mathrm{df}_2(\kappa,\kappa')$ and $\gamma_{\boldsymbol{M}} = q\,\mathrm{tr}[\boldsymbol{A G_A M G_A}]$ one can write these equivalences as:

$$\hat{\boldsymbol{G}}_{\boldsymbol{A}}\boldsymbol{M}\hat{\boldsymbol{G}}'_{\boldsymbol{A}} \simeq SS'\boldsymbol{G_A M G'_A} + SS'\boldsymbol{G_A A G'_A}\frac{\gamma_{\boldsymbol{M}}}{1-\gamma}, \tag{74}$$

$$\hat{\boldsymbol{T}}_{\boldsymbol{A}}\boldsymbol{M}\hat{\boldsymbol{T}}'_{\boldsymbol{A}} \simeq \boldsymbol{T_A M T'_A} + \kappa\kappa'\boldsymbol{G_A A G'_A}\frac{\gamma_{\boldsymbol{M}}}{1-\gamma}, \tag{75}$$

$$\hat{\boldsymbol{G}}_{\boldsymbol{A}}\boldsymbol{M}\hat{\boldsymbol{T}}'_{\boldsymbol{A}} \simeq S\boldsymbol{G_A M T'_A} - S\kappa'\boldsymbol{G_A A G'_A}\frac{\gamma_{\boldsymbol{M}}}{1-\gamma}, \tag{76}$$

$$\hat{\boldsymbol{T}}_{\boldsymbol{A}}\boldsymbol{M}\hat{\boldsymbol{G}}'_{\boldsymbol{A}} \simeq S'\boldsymbol{T_A M G'_A} - \kappa S'\boldsymbol{G_A A G'_A}\frac{\gamma_{\boldsymbol{M}}}{1-\gamma}. \tag{77}$$

D.2. Sanity Check of Two Point Functions

We now take $\boldsymbol{A} = \boldsymbol{\Sigma}, \boldsymbol{B} = \boldsymbol{W}$ for $\boldsymbol{W}$ a white Wishart and consider the matrix $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma} * \boldsymbol{W}$. In the case where $\boldsymbol{M} = \mathbf{I}$, the last two deterministic equivalences are equal. Further, by taking $\lambda = \lambda'$ so that $\kappa = \kappa'$ we get:

$$
\begin{aligned}
\hat{\mathbf{\Sigma}}(\hat{\mathbf{\Sigma}}+\lambda)^{-2} &\simeq S\mathbf{\Sigma}(\mathbf{\Sigma}+\kappa)^{-2} - S\kappa\mathbf{\Sigma}(\mathbf{\Sigma}+\kappa)^{-2}\frac{q\,\mathrm{tr}[\mathbf{\Sigma}(\mathbf{\Sigma}+\kappa)^{-2}]}{1-\gamma} \\
&= \mathbf{\Sigma}(\mathbf{\Sigma}+\kappa)^{-2}S\left[1-\frac{-q\kappa\mathrm{df}_1'}{1-\gamma}\right] \\
&= \frac{1}{1-\gamma}\mathbf{\Sigma}(\mathbf{\Sigma}+\kappa)^{-2}S\left[1-\gamma+q(\mathrm{df}_1-\mathrm{df}_2)\right] \\
&= \frac{1}{1-\gamma}\mathbf{\Sigma}(\mathbf{\Sigma}+\kappa)^{-2} \\
&= \frac{d\kappa}{d\lambda}\mathbf{\Sigma}(\mathbf{\Sigma}+\kappa)^{-2}.
\end{aligned} \tag{78}
$$

So we see that the the two-point result yields the same result as would be obtained through differentiation. This also extends to λ, λ' not equal. Taking $\mathrm{df}_2(\lambda,\lambda') = \mathrm{tr}[\mathbf{\Sigma}^2(\mathbf{\Sigma}+\lambda)^{-1}(\mathbf{\Sigma}+\lambda')^{-1}]$ and $\gamma = \frac{D}{P}\mathrm{df}_2(\lambda,\lambda')$, we have:

$$
\begin{aligned}
&\hat{\mathbf{\Sigma}}(\hat{\mathbf{\Sigma}}+\lambda)^{-1}(\hat{\mathbf{\Sigma}}+\lambda')^{-1} \\
&\simeq \mathbf{\Sigma}(\mathbf{\Sigma}+\kappa)^{-1}(\mathbf{\Sigma}+\kappa')^{-1}S\left(1-\kappa'\frac{q\,\mathrm{tr}[\mathbf{\Sigma}(\mathbf{\Sigma}+\kappa)^{-1}(\mathbf{\Sigma}+\kappa')^{-1}]}{1-\gamma}\right) \\
&= \frac{1}{1-\gamma}\mathbf{\Sigma}(\mathbf{\Sigma}+\kappa)^{-1}(\mathbf{\Sigma}+\kappa')^{-1}S(1-\gamma+q(\mathrm{df}_1-\mathrm{df}_2)) \\
&= \frac{1}{1-\gamma}\mathbf{\Sigma}(\mathbf{\Sigma}+\kappa)^{-1}(\mathbf{\Sigma}+\kappa')^{-1}.
\end{aligned} \tag{79}
$$

Here we have used that $\gamma = q\mathrm{df}_2$ and $S = (1-q\mathrm{df}_1)^{-1}$. This equation will be useful in treating various SGD kernel-related quantities.

Appendix E. Application I: Linear Regression

As a warm up, we first consider linear regression without random features. In this setting, one neglects $\boldsymbol{F}\boldsymbol{F}^\top$ by replacing it with the identity matrix. The dynamics in this setting are then:

$$
\hat{R}_t = \underbrace{\bar{\boldsymbol{w}}^\top e^{-2t\hat{\mathbf{\Sigma}}}\hat{\mathbf{\Sigma}}\bar{\boldsymbol{w}}}_{\hat{\mathcal{F}}(t)} + \chi\int_0^t \underbrace{\mathrm{Tr}[e^{-2(t-s)\hat{\mathbf{\Sigma}}}\hat{\mathbf{\Sigma}}^2]}_{\hat{\mathcal{K}}(t-s)}\hat{R}_s ds, \tag{80}
$$

$$
R_t = \underbrace{\bar{\boldsymbol{w}}^\top e^{-t\hat{\mathbf{\Sigma}}}\mathbf{\Sigma}e^{-t\hat{\mathbf{\Sigma}}}\bar{\boldsymbol{w}}}_{\mathcal{F}(t)} + \chi\int_0^t \underbrace{\mathrm{Tr}[e^{-2(t-s)\hat{\mathbf{\Sigma}}}\hat{\mathbf{\Sigma}}\mathbf{\Sigma}]}_{\mathcal{K}(t-s)}\hat{R}_s ds. \tag{81}
$$

E.1. Gradient Flow Term

The generalization error in Fourier space can then directly be obtained via the two-point master equation (69) with $\boldsymbol{M} = \boldsymbol{A} = \mathbf{\Sigma}$ to obtain

$$
\mathcal{F}(\omega,\omega') = \frac{SS'}{1-\gamma(i\omega_1, i\omega_1')}\bar{\boldsymbol{w}}^\top(i\omega_1+\mathbf{\Sigma})^{-1}\mathbf{\Sigma}(i\omega_1'+\mathbf{\Sigma})^{-1}\bar{\boldsymbol{w}}. \tag{82}
$$

Here, because $\hat{\mathbf{\Sigma}} = \mathbf{\Sigma} * \boldsymbol{W}$ for $\boldsymbol{W}$ a white Wishart, the renormalization of each frequency is given the multiplication of the S-transform of a white Wishart. This is found in the standard literature, see e.g. [26]. Our notation convention in this section is thus:

$$\begin{aligned} \mathrm{df}_1 &\equiv \mathrm{df}^1_{\hat{\mathbf{\Sigma}}}(\omega) \simeq \mathrm{df}^1_{\mathbf{\Sigma}}(\omega_1), \quad \mathrm{df}'_1 \equiv \mathrm{df}^1_{\hat{\mathbf{\Sigma}}}(\omega') \simeq \mathrm{df}^1_{\mathbf{\Sigma}}(\omega'_1) \\ S &\equiv \frac{1}{1-\frac{D}{P}\mathrm{df}_1}, \quad S' \equiv \frac{1}{1-\frac{D}{P}\mathrm{df}'_1}, \\ \omega_1 &\equiv S\omega, \quad \omega'_1 \equiv S'\omega', \\ \mathrm{df}_2 &\equiv \mathrm{tr}\left[\mathbf{\Sigma}^2(i\omega_1+\mathbf{\Sigma})^{-1}(i\omega'_1+\mathbf{\Sigma})^{-1}\right]. \\ \gamma &\equiv \frac{D}{P}\mathrm{df}_2. \end{aligned} \tag{83}$$

The empirical forcing term can be handled with a single frequency Fourier transform. Indeed it is much more convenient to do so when dealing with the SGD effects of the next section. We can write it as:

$$\hat{\mathcal{F}}(t) = \int_\omega e^{2i\omega t}\bar{\boldsymbol{w}}^\top \hat{\mathbf{\Sigma}}(\hat{\mathbf{\Sigma}}+i\omega)^{-1} \simeq \int_\omega e^{2i\omega t}\bar{\boldsymbol{w}}^\top \mathbf{\Sigma}(\mathbf{\Sigma}+i\omega_1)^{-1}. \tag{84}$$

In the last equality, we have applied a strong (one point) deterministic equivalence.

We consider instead the bi-frequency transformation of the empirical loss in Appendix G.1, and see that it gives a dynamical analogue of the generalized cross-validation (GCV) in the gradient flow limit.

E.2. SGD Kernel Term

In order to correctly treat the convolution of the kernel term with the empirical risk, it is much easier to work in single frequency Fourier space. We now evaluate both the train and the test kernel.

$$\mathcal{K}_t = \mathrm{Tr}[e^{-2t\hat{\mathbf{\Sigma}}}\hat{\mathbf{\Sigma}}\mathbf{\Sigma}] = \int_\omega e^{2i\omega t}\mathrm{Tr}[\mathbf{\Sigma}\hat{\mathbf{\Sigma}}(\hat{\mathbf{\Sigma}}+i\omega)^{-1}]. \tag{85}$$

We can apply one-point deterministic equivalent to obtain:

$$\mathrm{Tr}[\mathbf{\Sigma}\hat{\mathbf{\Sigma}}(\hat{\mathbf{\Sigma}}+i\omega)^{-1}] \simeq \mathrm{Tr}[\mathbf{\Sigma}^2(\mathbf{\Sigma}+i\omega_1)^{-1}] = \mathrm{Tr}[\mathbf{\Sigma}] - i\omega_1\mathrm{Tr}[\mathbf{\Sigma}(\mathbf{\Sigma}+i\omega_1)^{-1}] \tag{86}$$

Similarly, for the train kernel, we have

$$\hat{\mathcal{K}}_t = \mathrm{Tr}[e^{-2t\hat{\mathbf{\Sigma}}}\hat{\mathbf{\Sigma}}^2] = \int_\omega e^{2i\omega t}\mathrm{Tr}[\hat{\mathbf{\Sigma}}^2(\hat{\mathbf{\Sigma}}+i\omega)^{-1}]. \tag{87}$$

In Fourier space, this resolvent is given by:

$$\begin{aligned} \mathrm{Tr}[\hat{\mathbf{\Sigma}}^2(\hat{\mathbf{\Sigma}}+i\omega)^{-1}] &= \mathrm{Tr}[\hat{\mathbf{\Sigma}}] - i\omega\mathrm{Tr}[\hat{\mathbf{\Sigma}}(\hat{\mathbf{\Sigma}}+i\omega)^{-1}] \\ &\simeq \mathrm{Tr}[\mathbf{\Sigma}] - i\omega\mathrm{Tr}[\mathbf{\Sigma}(\mathbf{\Sigma}+i\omega_1)^{-1}] \\ &= \mathrm{Tr}[\mathbf{\Sigma}^2(\mathbf{\Sigma}+i\omega_1)^{-1}] + i(\omega_1-\omega)\mathrm{Tr}[\mathbf{\Sigma}(\mathbf{\Sigma}+i\omega_1)^{-1}]. \\ &= \mathcal{K}(\omega) + \frac{i\omega_1}{P}\mathrm{Tr}[\mathbf{\Sigma}(\mathbf{\Sigma}+i\omega_1)^{-1}]^2 \end{aligned} \tag{88}$$

Where we have used that $i(\omega_1 - \omega) = i\omega(S-1) = i\omega_1\frac{D}{P}\mathrm{df}_1$. We thus see that the deterministic equivalent for the train kernel is the same as the equivalent for the test kernel, but with an extra additive term. In the large t, or equivalently $\omega \to 0$ limit, we have that this additional term is always *non-negative*.

Given the exact expression for the test risk, we get that the additive SGD contributions to the train and test risk are given by

$$\begin{aligned}\hat{R}(t) &= \int_\omega e^{2i\omega t} \frac{\hat{\mathcal{F}}(\omega)}{1-\hat{\mathcal{K}}(\omega)} \\ &= \int_\omega e^{2i\omega t} \frac{\bar{\boldsymbol{w}}^\top \boldsymbol{\Sigma}(\boldsymbol{\Sigma}+i\omega_1)^{-1}\bar{\boldsymbol{w}}}{1-\chi\mathrm{Tr}[\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+i\omega_1)^{-1}]-\chi i\omega_1(D\mathrm{df}_1)^2/P}\end{aligned} \tag{89}$$

and

$$\begin{aligned}R(t) &= \int_{\omega,\omega'} e^{i(\omega+\omega')t}\mathcal{F}(\omega,\omega') + \int_\omega e^{2i\omega t}\frac{\mathcal{K}(\omega)\hat{\mathcal{F}}(\omega)}{1-\hat{\mathcal{K}}(\omega)} \\ &= \int_{\omega,\omega'} \frac{e^{i(\omega+\omega')t}SS'}{1-\gamma(i\omega_1,i\omega_1')}\bar{\boldsymbol{w}}^\top(i\omega_1+\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}(i\omega_1'+\boldsymbol{\Sigma})^{-1}\bar{\boldsymbol{w}} \\ &\quad + \int_\omega e^{2i\omega t}\frac{\chi\mathrm{Tr}[\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+i\omega_1)^{-1}]}{1-\chi\mathrm{Tr}[\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+i\omega_1)^{-1}]-\chi i\omega_1(D\mathrm{df}_1)^2/P} \\ &\qquad \times \bar{\boldsymbol{w}}^\top\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+i\omega_1)^{-1}\bar{\boldsymbol{w}},\end{aligned} \tag{90}$$

respectively. In Appendix G.2, we compute the SGD terms in terms of double frequency Fourier transforms and two-point equivalents.

E.3. Covariate Shift

In this section, we show how the two-point deterministic equivalences allow for direct calculation of the train and test risks under a shift of covariates from $\boldsymbol{\Sigma}$ to $\boldsymbol{\Sigma}'$. There, the forcing term changes to

$$\begin{aligned}\mathcal{F}_{OOD}(\omega,\omega') &= SS'\bar{\boldsymbol{w}}^\top(i\omega_1+\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}'(i\omega_1'+\boldsymbol{\Sigma})^{-1}\bar{\boldsymbol{w}} \\ &\quad + SS'\frac{\gamma_{\boldsymbol{\Sigma}'}}{1-\gamma}\bar{\boldsymbol{w}}^\top(i\omega_1+\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}(i\omega_1'+\boldsymbol{\Sigma})^{-1}\bar{\boldsymbol{w}}.\end{aligned} \tag{91}$$

In the above equation,

$$\gamma_{\boldsymbol{\Sigma}'}(i\omega_1,i\omega_1') \equiv \frac{D}{P}\,\mathrm{tr}\left[\boldsymbol{\Sigma}(i\omega_1+\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}'(i\omega_1'+\boldsymbol{\Sigma})^{-1}\right]. \tag{92}$$

The (single frequency) SGD test kernel term also changes to:

$$\mathcal{K}'(\omega) \simeq \mathrm{Tr}[\boldsymbol{\Sigma}\boldsymbol{\Sigma}'(\boldsymbol{\Sigma}+i\omega_1)^{-1}] = \mathrm{Tr}[\boldsymbol{\Sigma}'] - i\omega_1 D\mathrm{df}_1. \tag{93}$$

In the static limit, we have $\kappa = \lim_{\omega=0} i\omega_1$. We also have explicitly written out the noise term rather than including it as a mode at infinity. Then, the distribution-shifted gradient flow term becomes:

$$\begin{aligned}E_{OOD}^{\boldsymbol{\Sigma}',\bar{\boldsymbol{w}}} &\simeq \kappa^2\left[\bar{\boldsymbol{w}}^\top(\boldsymbol{\Sigma}+\kappa)^{-1}\boldsymbol{\Sigma}'(\boldsymbol{\Sigma}+\kappa)^{-1}\bar{\boldsymbol{w}} + \bar{\boldsymbol{w}}^\top\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+\kappa)^{-2}\bar{\boldsymbol{w}}\frac{\gamma'}{1-\gamma}\right] \\ &\quad + \sigma_\epsilon^2\frac{\gamma'}{1-\gamma}\end{aligned} \tag{94}$$

This recovers the covariate-shifted results of [3, 9, 25]. We note that one can also easily accommodate target shifts (*i.e.*, from $\bar{\boldsymbol{w}}$ to $\bar{\boldsymbol{w}}'$) in this formalism, but doing so does not require two-point equivalents [25].

Appendix F. Application II: Random Feature Regression

We now return to our original model, with random features added. When data averaging, ω will be renormalized by the S-transform of a white Wishart $S_{\boldsymbol{W}}$ as before. When averaging over features, it will further be renormalized by the S-transform $S_{\boldsymbol{F}\boldsymbol{F}^\top}$. Here, $\boldsymbol{F}\boldsymbol{F}^\top$ is a white Wishart matrix with parameter $q = D/N$. Our notation will thus be:

$$\begin{aligned}
\mathrm{df}_1 &\equiv \mathrm{df}^1_{\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}}(\omega), \quad \mathrm{df}'_1 \equiv \mathrm{df}^1_{\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}}(\omega'),\\
S_{\boldsymbol{W}} &\equiv \frac{1}{1-\frac{D}{P}\mathrm{df}_1}, \quad S'_{\boldsymbol{W}} \equiv \frac{1}{1-\frac{D}{P}\mathrm{df}'_1},\\
S_{\boldsymbol{F}\boldsymbol{F}^\top} &\equiv \frac{1}{1-\frac{D}{N}\mathrm{df}_1}, \quad S'_{\boldsymbol{F}\boldsymbol{F}^\top} \equiv \frac{1}{1-\frac{D}{N}\mathrm{df}'_1},\\
S &\equiv S_{\boldsymbol{W}} S_{\boldsymbol{F}\boldsymbol{F}^\top}, \quad S' \equiv S'_{\boldsymbol{W}} S'_{\boldsymbol{F}\boldsymbol{F}^\top},\\
\omega_1 &\equiv S_{\boldsymbol{W}}\omega, \quad \omega'_1 \equiv S'_{\boldsymbol{W}}\omega',\\
\omega_2 &\equiv S_{\boldsymbol{F}\boldsymbol{F}^\top}\omega_1 = S\omega, \quad \omega'_2 \equiv S'_{\boldsymbol{F}\boldsymbol{F}^\top}\omega'_1 = S'\omega',\\
\mathrm{df}_2 &\equiv \mathrm{tr}[\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+i\omega_2)^{-1}(\boldsymbol{\Sigma}+i\omega'_2)^{-1}].
\end{aligned} \tag{95}$$

Then, the weak deterministic equivalents for df_1 and df'_1 can be written as

$$\begin{aligned}
\mathrm{df}_1 &\equiv \mathrm{df}^1_{\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}}(\omega) \simeq \mathrm{df}^1_{\boldsymbol{F}\boldsymbol{F}^\top\boldsymbol{\Sigma}}(\omega_1) \simeq \mathrm{df}^1_{\boldsymbol{\Sigma}}(\omega_2),\\
\mathrm{df}'_1 &\equiv \mathrm{df}^1_{\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}}(\omega') \simeq \mathrm{df}^1_{\boldsymbol{F}\boldsymbol{F}^\top\boldsymbol{\Sigma}}(\omega'_1) \simeq \mathrm{df}^1_{\boldsymbol{\Sigma}}(\omega'_2),
\end{aligned} \tag{96}$$

where we average first over the randomness in $\hat{\boldsymbol{\Sigma}}$ and then $\boldsymbol{F}\boldsymbol{F}^\top$. As stated before, in contrast to [4], we choose to divide by N instead of D in defining the features. This will lead to a more natural gradient flow dynamics. In the convention of [4], one would have to re-scale $\eta \to \eta D/N$ to get correct dynamics. Either way, the large time limit will agree with agree with the (ridgeless) random feature results quoted in that work.

We will also make frequent use of the **push-through identity**:

$$\boldsymbol{A}(\boldsymbol{B}\boldsymbol{A}+\lambda)^{-1} = (\boldsymbol{A}\boldsymbol{B}+\lambda)^{-1}\boldsymbol{A}. \tag{97}$$

F.1. Gradient Flow Term

By writing $\hat{\boldsymbol{\Sigma}} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{W}\boldsymbol{\Sigma}^{1/2}$ for $\boldsymbol{W}$ a white Wishart matrix and applying the push-through identity (97) to the original random feature forcing equation (37), we see that we need to evaluate:

$$\begin{aligned}
\mathcal{F}(i\omega, i\omega') = \bar{\boldsymbol{w}}^\top\boldsymbol{\Sigma}^{1/2}&(i\omega + \boldsymbol{W}\boldsymbol{\Sigma}^{1/2}\boldsymbol{F}\boldsymbol{F}^\top\boldsymbol{\Sigma}^{1/2})^{-1}\\
&\times (i\omega' + \boldsymbol{\Sigma}^{1/2}\boldsymbol{F}\boldsymbol{F}^\top\boldsymbol{\Sigma}^{1/2}\boldsymbol{W})^{-1}\,\boldsymbol{\Sigma}^{1/2}\bar{\boldsymbol{w}}.
\end{aligned} \tag{98}$$

We do this in two steps, first integrating over data $\boldsymbol{W}$ and then over random features $\boldsymbol{F}\boldsymbol{F}^\top$.

F.1.1. Integrating Over Data

Pushing through $\boldsymbol{\Sigma}^{1/2}$ on both sides, we apply the two-point equivalence (71) with $\boldsymbol{M} = \mathbf{I}$ and $\boldsymbol{A} = \boldsymbol{\Sigma}^{1/2}\boldsymbol{F}\boldsymbol{F}^\top\boldsymbol{\Sigma}^{1/2}$ to obtain:

$$\mathcal{F}(i\omega, i\omega') \simeq \frac{S_{\boldsymbol{W}} S'_{\boldsymbol{W}}}{1-\gamma_1}\bar{\boldsymbol{w}}^\top(i\omega_1 + \boldsymbol{\Sigma}\boldsymbol{F}\boldsymbol{F}^\top)^{-1}\boldsymbol{\Sigma}(i\omega'_1 + \boldsymbol{F}\boldsymbol{F}^\top\boldsymbol{\Sigma})^{-1}\bar{\boldsymbol{w}} \tag{99}$$

for

$$\gamma_1 \equiv \frac{D}{P}\,\mathrm{tr}\left[(\boldsymbol{\Sigma}\boldsymbol{F}\boldsymbol{F}^\top)^2(i\omega_1 + \boldsymbol{\Sigma}\boldsymbol{F}\boldsymbol{F}^\top)^{-1}(i\omega'_1 + \boldsymbol{\Sigma}\boldsymbol{F}\boldsymbol{F}^\top)^{-1}\right]. \tag{100}$$

F.1.2. INTEGRATING OVER FEATURES

We will evaluate γ_1 separately since it concentrates. We apply Equation (72) for a *Gram Wishart* with $q = D/N, q^{-1} = N/D$. Here, $\boldsymbol{M} = \mathbf{I}$, $\boldsymbol{A} = \boldsymbol{\Sigma}$, $B = \boldsymbol{F}\boldsymbol{F}^\top$.

$$\gamma_1 \simeq \frac{D}{P}\mathrm{df}_2 + \frac{D}{P}\frac{D}{N}(i\omega_2)(i\omega_2')\frac{\mathrm{tr}[(i\omega_2 + \boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}(i\omega_2' + \boldsymbol{\Sigma})^{-1}]^2}{1 - \frac{D}{N}\mathrm{df}_2}. \tag{101}$$

We now apply (68) with $\boldsymbol{M} = \mathbf{I}$, $\boldsymbol{A} = \boldsymbol{\Sigma}$, and $\boldsymbol{B} = \boldsymbol{F}\boldsymbol{F}^\top$ to get:

$$\begin{aligned}\mathcal{F}(i\omega, i\omega') \simeq \frac{SS'}{1-\gamma_1}\Bigg[&\bar{\boldsymbol{w}}^\top(i\omega_2 + \boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}(i\omega_2' + \boldsymbol{\Sigma})^{-1}\bar{\boldsymbol{w}} \\ &+ \bar{\boldsymbol{w}}^\top(i\omega_2 + \boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}^2(i\omega_2' + \boldsymbol{\Sigma})^{-1}\bar{\boldsymbol{w}} \\ &\times \frac{\frac{D}{N}\,\mathrm{tr}[(i\omega_2 + \boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}(i\omega_2' + \boldsymbol{\Sigma})^{-1}]}{1 - \frac{D}{N}\mathrm{df}_2(i\omega_2, i\omega_2')}\Bigg].\end{aligned} \tag{102}$$

This recovers the result of [8].

We now evaluate the training loss. Because of the push-through identity, even in the random feature setting we can characterize this in single-frequency space using a one-point deterministic equivalent:

$$\begin{aligned}\hat{\mathcal{F}}(t) &= \bar{\boldsymbol{w}}^\top e^{-t\hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top}\hat{\boldsymbol{\Sigma}}e^{-t\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}}\bar{\boldsymbol{w}} = \bar{\boldsymbol{w}}^\top\hat{\boldsymbol{\Sigma}}e^{-2t\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}}\bar{\boldsymbol{w}} \\ &= \int_\omega e^{2i\omega t}\bar{\boldsymbol{w}}^\top\hat{\boldsymbol{\Sigma}}(\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}} + i\omega)^{-1}\bar{\boldsymbol{w}} \\ &\simeq \int_\omega e^{2i\omega t}S_{\boldsymbol{F}\boldsymbol{F}^\top}\bar{\boldsymbol{w}}^\top\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + i\omega_2)^{-1}\bar{\boldsymbol{w}}.\end{aligned} \tag{103}$$

In the last line we have applied the one-point deterministic equivalence twice, over $\boldsymbol{F}\boldsymbol{F}^\top$ and over $\boldsymbol{W}$.

F.2. SGD Kernel Term

Here, we again apply a one-point deterministic equivalence to characterize the SGD kernel in single-frequency space. For the test kernel, we have:

$$\mathcal{K}(t) = \int_\omega e^{2i\omega t}\mathrm{Tr}[\boldsymbol{\Sigma}\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top(\hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top + i\omega)^{-1}] \tag{104}$$

We apply the following two deterministic equivalences:

$$\begin{aligned}&\mathrm{Tr}[\boldsymbol{\Sigma}\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top(\hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top + i\omega)^{-1}] \\ &\quad\simeq \mathrm{Tr}[\boldsymbol{\Sigma}\boldsymbol{F}\boldsymbol{F}^\top\boldsymbol{\Sigma}\boldsymbol{F}\boldsymbol{F}^\top(\boldsymbol{\Sigma}\boldsymbol{F}\boldsymbol{F}^\top + i\omega_1)^{-1}] \\ &\quad= \mathrm{Tr}[\boldsymbol{\Sigma}\boldsymbol{F}\boldsymbol{F}^\top] - i\omega_1\mathrm{Tr}[\boldsymbol{\Sigma}\boldsymbol{F}\boldsymbol{F}^\top(\boldsymbol{\Sigma}\boldsymbol{F}\boldsymbol{F}^\top + i\omega_1)^{-1}] \\ &\quad\simeq \mathrm{Tr}[\boldsymbol{\Sigma}] - i\omega_1\mathrm{Tr}[\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + i\omega_2)^{-1}]. \\ &\quad\simeq \mathrm{Tr}[\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + i\omega_2)^{-1}] + i(\omega_2 - \omega_1)\mathrm{Tr}[\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + i\omega_2)^{-1}] \\ &\quad\simeq \mathrm{Tr}[\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma} + i\omega_2)^{-1}] + \frac{i\omega_2}{N}\mathrm{Tr}[\boldsymbol{\Sigma}(\boldsymbol{\Sigma} + i\omega_2)^{-1}]^2.\end{aligned} \tag{105}$$

Similarly, for the train kernel, we have:

$$\hat{\mathcal{K}}(t) = \int_\omega e^{2i\omega t}\mathrm{Tr}[\hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top(\hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top + i\omega)^{-1}]. \tag{106}$$

Here, we apply two deterministic equivalences again:

$$\begin{aligned}
&\mathrm{Tr}[\hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top(\hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top+i\omega)^{-1}] \\
&\quad= \mathrm{Tr}[\hat{\boldsymbol{\Sigma}}\boldsymbol{F}\boldsymbol{F}^\top] - i\omega\mathrm{Tr}[\boldsymbol{\Sigma}\boldsymbol{F}\boldsymbol{F}^\top(\boldsymbol{\Sigma}\boldsymbol{F}\boldsymbol{F}^\top+i\omega)^{-1}] \\
&\quad\simeq \mathrm{Tr}[\boldsymbol{\Sigma}] - i\omega\mathrm{Tr}[\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+i\omega_2)^{-1}]. \\
&\quad= \mathrm{Tr}[\boldsymbol{\Sigma}^2(\boldsymbol{\Sigma}+i\omega_2)^{-1}] + i(\omega_2-\omega)\mathrm{Tr}[\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+i\omega_2)^{-1}] \\
&\quad= \mathcal{K}(\omega) + i(\omega_1-\omega)\mathrm{Tr}[\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+i\omega_2)^{-1}] \\
&\quad= \mathcal{K}(\omega) + \frac{i\omega_1}{P}\mathrm{Tr}[\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+i\omega_2)^{-1}]^2.
\end{aligned} \tag{107}$$

All together, the empirical and population risk are then

$$\begin{aligned}
\hat{R}(t) &= \int_\omega e^{2i\omega t}\frac{S_{\boldsymbol{F}\boldsymbol{F}^\top}\bar{\boldsymbol{w}}^\top\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+i\omega_2)^{-1}\bar{\boldsymbol{w}}}{1-\chi\mathrm{Tr}[\boldsymbol{\Sigma}]+i\omega\chi D\mathrm{df}_1}, \\
R(t) &= \int_{\omega,\omega'} e^{i(\omega+\omega')t}\mathcal{F}(\omega,\omega') \\
&\quad+ \int_\omega e^{2i\omega t}\frac{\chi\mathrm{Tr}[\boldsymbol{\Sigma}]-i\omega_1\chi D\mathrm{df}_1}{1-\chi\mathrm{Tr}[\boldsymbol{\Sigma}]+i\omega\chi D\mathrm{df}_1}S_{\boldsymbol{F}\boldsymbol{F}^\top}\bar{\boldsymbol{w}}^\top\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+i\omega_2)^{-1}\bar{\boldsymbol{w}}.
\end{aligned} \tag{108}$$

Here, $\mathcal{F}(\omega,\omega')$ is the forcing term reported in Equation (102).

Appendix G. Double-Frequency Equivalents for the Empirical Forcing and SGD Kernel Terms

In this section we match the double-frequency treatment of the population forcing term. For the empirical forcing terms, the bi-frequency picture yields an analogue of the generalized cross-validation (GCV) procedure in each frequency mode. The utility of the bi-frequency picture for the kernel forcing terms is less clear, but we report it anyway for completeness.

G.1. Linear Regression Forcing Term

We can write the empirical loss under gradient flow in terms of two time variables as:

$$\hat{\mathcal{F}}(t,t') \equiv \Delta\boldsymbol{w}_t^\top\hat{\boldsymbol{\Sigma}}\Delta\boldsymbol{w}_{t'} = \bar{\boldsymbol{w}}^\top e^{-t\hat{\boldsymbol{\Sigma}}}\hat{\boldsymbol{\Sigma}}e^{-t'\hat{\boldsymbol{\Sigma}}}\bar{\boldsymbol{w}}. \tag{109}$$

Then, Fourier transforming in each variable separately and restricting to the $t = t'$ diagonal yields:

$$\hat{\mathcal{F}}(t) = \int_{\omega,\omega'} e^{it(\omega+\omega')}\underbrace{\bar{\boldsymbol{w}}^\top\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Sigma}}+i\omega)^{-1}(\hat{\boldsymbol{\Sigma}}+i\omega')^{-1}\bar{\boldsymbol{w}}}_{\hat{\mathcal{F}}(\omega,\omega')}. \tag{110}$$

We now apply the deterministic equivalence (79) to obtain:

$$\hat{\mathcal{F}}(\omega,\omega') \simeq \frac{1}{1-\gamma}\bar{\boldsymbol{w}}^\top\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+i\omega_1)^{-1}(\boldsymbol{\Sigma}+i\omega_1')^{-1}\bar{\boldsymbol{w}}. \tag{111}$$

G.2. SGD in Linear Regression

Here, to match the gradient flow term, we apply two Fourier transforms to the kernel and apply the two-point deterministic equivalence. Namely, we consider extending $\mathcal{K}, \hat{\mathcal{K}}$ to:

$$\mathcal{K}(t,t') \equiv \mathrm{Tr}[e^{-t\hat{\boldsymbol{\Sigma}}}\hat{\boldsymbol{\Sigma}}e^{-t'\hat{\boldsymbol{\Sigma}}}\boldsymbol{\Sigma}], \quad \hat{\mathcal{K}}(t,t') \equiv \mathrm{Tr}[e^{-t\hat{\boldsymbol{\Sigma}}}\hat{\boldsymbol{\Sigma}}e^{-t'\hat{\boldsymbol{\Sigma}}}\hat{\boldsymbol{\Sigma}}]. \tag{112}$$

We then perform Fourier transforms separately in t, t' to obtain:

$$\begin{aligned}\mathcal{K}(t) &= D \int_{\omega,\omega'} e^{it(\omega+\omega')} \underbrace{\mathrm{tr}[\boldsymbol{\Sigma}(\hat{\boldsymbol{\Sigma}}+i\omega)^{-1}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Sigma}}+i\omega')^{-1}]}_{\mathcal{K}(\omega,\omega')},\\ \hat{\mathcal{K}}(t) &= D \int_{\omega,\omega'} e^{it(\omega+\omega')} \underbrace{\mathrm{tr}[\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Sigma}}+i\omega)^{-1}\hat{\boldsymbol{\Sigma}}(\hat{\boldsymbol{\Sigma}}+i\omega')^{-1}]}_{\hat{\mathcal{K}}(\omega,\omega')}.\end{aligned} \tag{113}$$

By applying the two-point master formulas in Appendix D.1, and specifically (79), we obtain the equivalent for the test kernel:

$$\mathcal{K}(\omega,\omega') \simeq \frac{D\mathrm{df}_2}{1-\gamma}. \tag{114}$$

Similarly the train kernel can be written as:

$$\hat{\mathcal{K}}(\omega,\omega') \simeq D\mathrm{df}_2\left(1-(i\omega_1)(i\omega_1')\frac{D}{P}\frac{\mathrm{tr}[\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+i\omega_1)^{-1}(\boldsymbol{\Sigma}+i\omega_1')^{-1}]^2}{1-\gamma}\right). \tag{115}$$

G.3. Linear Random Features Forcing Term

The empirical loss under gradient flow in bi-frequency space is given by

$$\hat{\mathcal{F}}(t) = \int_{\omega,\omega'} e^{it(\omega+\omega')}\bar{\boldsymbol{w}}^\top\hat{\boldsymbol{\Sigma}}(\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}+i\omega)^{-1}(\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}+i\omega')^{-1}\bar{\boldsymbol{w}}. \tag{116}$$

We now apply the deterministic equivalence (79) to obtain:

$$\hat{\mathcal{F}}(\omega,\omega') \simeq \frac{1}{1-\gamma_1}\bar{\boldsymbol{w}}^\top\boldsymbol{\Sigma}(\boldsymbol{\Sigma}\boldsymbol{F}\boldsymbol{F}^\top+i\omega_1)^{-1}(\boldsymbol{\Sigma}\boldsymbol{F}\boldsymbol{F}^\top+i\omega_1')^{-1}\bar{\boldsymbol{w}}. \tag{117}$$

Finally, we apply the deterministic equivalence (69) over $\boldsymbol{F}\boldsymbol{F}^\top$ to yield:

$$\begin{aligned}\hat{\mathcal{F}}(\omega,\omega') &\simeq \frac{S_{\boldsymbol{F}\boldsymbol{F}^\top}S'_{\boldsymbol{F}\boldsymbol{F}^\top}}{1-\gamma_1}\Bigg[\bar{\boldsymbol{w}}^\top\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+i\omega_2)^{-1}(\boldsymbol{\Sigma}+i\omega_2')^{-1}\bar{\boldsymbol{w}}\\ &\qquad + \bar{\boldsymbol{w}}^\top(i\omega_2+\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}^2(i\omega_2'+\boldsymbol{\Sigma})^{-1}\bar{\boldsymbol{w}}\\ &\qquad\quad \times \frac{\frac{D}{N}\mathrm{tr}[(i\omega_2+\boldsymbol{\Sigma})^{-1}\boldsymbol{\Sigma}(i\omega_2'+\boldsymbol{\Sigma})^{-1}]}{1-\frac{D}{N}\mathrm{df}_2(i\omega_2,i\omega_2')}\Bigg]\\ &= \frac{\mathcal{F}(\omega,\omega')}{S_{\boldsymbol{W}}S'_{\boldsymbol{W}}}.\end{aligned} \tag{118}$$

This again yields a dynamical analogue of GCV, where we see that the empirical and population risks under gradient flow differ by a factor of $S_{\boldsymbol{W}}S'_{\boldsymbol{W}}$. In the $t\to\infty$ limit this requires the $S_{\boldsymbol{W}}^2$ obtained in [1, 4] for linear random features.

G.4. SGD in Linear Random Features

We now compute deterministic equivalents for the kernel term, again defining the bi-temporal kernels by:

$$\begin{aligned}\mathcal{K}(t,t') &\equiv \mathrm{Tr}[e^{-t\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}}\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}e^{-t'\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}}\boldsymbol{F}\boldsymbol{F}^\top\boldsymbol{\Sigma}],\\ \hat{\mathcal{K}}(t,t') &\equiv \mathrm{Tr}[e^{-t\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}}\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}e^{-t'\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}}\boldsymbol{F}\boldsymbol{F}^\top\hat{\boldsymbol{\Sigma}}].\end{aligned} \tag{119}$$

We then perform Fourier transforms separately in t, t' to obtain:

$$\begin{aligned} \mathcal{K}(t) &= D \int_{\omega,\omega'} e^{it(\omega+\omega')} \mathcal{K}(\omega,\omega'), \\ \hat{\mathcal{K}}(t) &= D \int_{\omega,\omega'} e^{it(\omega+\omega')} \hat{\mathcal{K}}(\omega,\omega') \end{aligned} \tag{120}$$

for

$$\begin{aligned} \mathcal{K}(\omega,\omega') &\equiv \mathrm{tr}[\boldsymbol{F}\boldsymbol{F}^\top \boldsymbol{\Sigma}(\boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}} + i\omega)^{-1} \boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}}(\boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}} + i\omega')^{-1}] \\ \hat{\mathcal{K}}(\omega,\omega') &\equiv \mathrm{tr}[\boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}}(\boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}} + i\omega)^{-1} \boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}}(\boldsymbol{F}\boldsymbol{F}^\top \hat{\boldsymbol{\Sigma}} + i\omega')^{-1}]. \end{aligned} \tag{121}$$

Applying the two-point equivalences to perform the data average, we get for the test kernel:

$$\mathcal{K}(\omega,\omega') \simeq \frac{D\mathrm{df}^2_{\boldsymbol{F}\boldsymbol{F}^\top\boldsymbol{\Sigma}}(i\omega_1, i\omega_1')}{1-\gamma_1(i\omega_1, i\omega_1')}. \tag{122}$$

We recognize again that the numerator and denominator depend only on $\gamma_1 \equiv \frac{D}{P}\mathrm{df}^2_{\boldsymbol{F}\boldsymbol{F}^\top\boldsymbol{\Sigma}}$, We have already computed equivalents for this in the prior section.

Similarly for the train kernel we get:

$$\begin{aligned} \hat{\mathcal{K}}(\omega,\omega') &\simeq D\mathrm{df}^2_{\boldsymbol{F}\boldsymbol{F}^\top\boldsymbol{\Sigma}}(i\omega_1, i\omega_1') \\ &\quad\times \left[1-(i\omega_1)(i\omega_1')\frac{D}{P}\frac{\mathrm{tr}[\boldsymbol{\Sigma}\boldsymbol{F}\boldsymbol{F}^\top(\boldsymbol{\Sigma}\boldsymbol{F}\boldsymbol{F}^\top + i\omega_1)^{-1}(\boldsymbol{\Sigma}\boldsymbol{F}\boldsymbol{F}^\top + i\omega_1')^{-1}]}{1-\gamma_1(i\omega_1, i\omega_1')}\right]. \end{aligned} \tag{123}$$

Again equivalents for $\mathrm{df}^2_{\boldsymbol{F}\boldsymbol{F}^\top\boldsymbol{\Sigma}}$ and γ_1 are already calculated. It remains to apply a final deterministic equivalence (79) over the features to the last term to get:

$$\begin{aligned} &\mathrm{tr}[\boldsymbol{\Sigma}\boldsymbol{F}\boldsymbol{F}^\top(\boldsymbol{\Sigma}\boldsymbol{F}\boldsymbol{F}^\top + i\omega_1)^{-1}(\boldsymbol{\Sigma}\boldsymbol{F}\boldsymbol{F}^\top + i\omega_1')^{-1}] \\ &\quad\simeq \frac{\mathrm{tr}[\boldsymbol{\Sigma}(\boldsymbol{\Sigma}+i\omega_2)^{-1}(\boldsymbol{\Sigma}+i\omega_2')^{-1}]}{1-\frac{D}{N}\mathrm{df}_2(i\omega_2, i\omega_2')}. \end{aligned} \tag{124}$$