Supplemental Material

Double Trouble in Double Descent:
Bias and Variance(s) in the Lazy Regime

Contents

1 Further analytical results

1.1 Asymptotic scalings

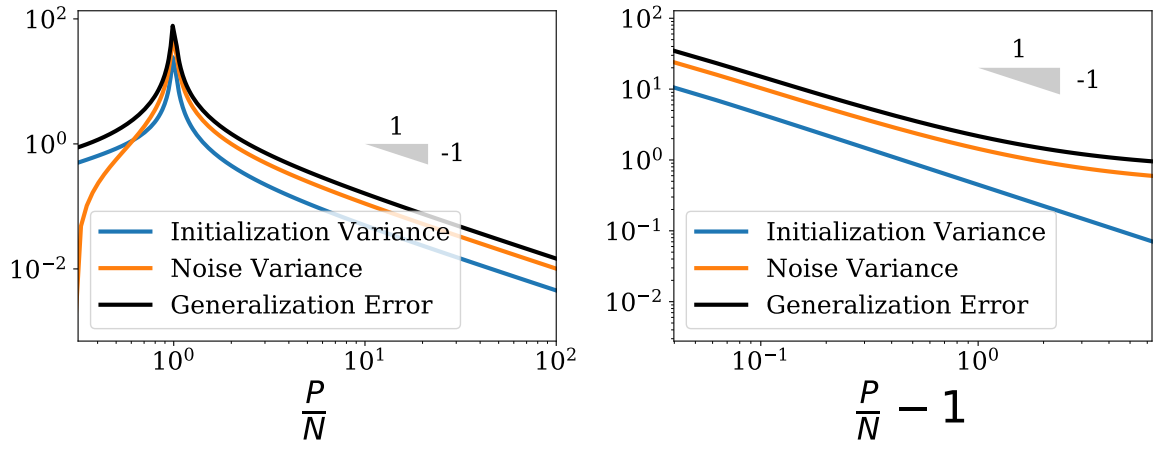

Figure 1: Log-log behaviour of the quantities of interest at **Left:** $P/N \to \infty$ and **Right**: $P \to N$ with $\lambda = 10^{-5}$, $N/D = 1$ and $\tau = 1$. In both cases, one observes an inverse scaling law.

Figure 1 (left) shows that the various terms entering the decomposition of the test error approach their asymptotic values at a rate $(P/N)^{-1}$. This scaling law is consistent with that found in [3] for real neural networks, where P is replaced by the width of the layers of the network. As for the divergence of the noise and initialization variances observed at the interpolation threshold, figure 1 (right) shows that they also follow an inverse power law $(P/N-1)^{-1}$ at vanishing regularization.

1.2 Divide and Conquer approach

As mentioned in the main text, another way to average the predictions of differently initialized learners is the *divide and conquer* approach [2]. In this framework, the data set is divided into K splits of size N/K. Each of the K differently initalized learner is trained on a distinct split. This approach is extremely useful for kernel learning [5], where the computational burden is in the inversion of the Gram matrix which is of size $N \times N$. In the random projection approach considered here, it does not offer any computational gain, however it is interesting how it affects the test error.
Within our framework, the test error can easily be calculated as:

$$\underset{\{\boldsymbol{\Theta}^{(k)}\},\boldsymbol{X},\boldsymbol{\varepsilon}}{\mathbb{E}}[\mathcal{R}_{\rm RF}] = F^2\left(1-2\Psi_1^v\right)+\frac{1}{K}\left(F^2\Psi_2^v+\tau^2\Psi_3^v\right)+\left(1-\frac{1}{K}\right)F^2\Psi_2^d, \tag{1}$$

where the effective number of data points which enters this formula is $N_{\rm eff} = N/K$ due to the splitting of the training set.
Comparing the previous expression with that obtained for ensembling (24) is instructive: here, increasing K replaces the *vanilla* terms Ψ_2^v, Ψ_3^v by the *divide and conquer* term Ψ_2^d. This shows that divide and conquer has a *denoising* effect: at $K \to \infty$, the effect of the additive noise on the labels is completely suppressed. This was not the case for ensembling. The price to pay is that $N_{\rm eff}$ decreases, hence one is shifted to the underparametrized regime.
In Figure 2, we see that the kernel limit error of the divide and conquer approach, i.e. the asymptotic value of the error at $P/N \to \infty$, is different from the usual kernel limit error, since the effective dataset is two times smaller at $K = 2$. The denoising effect of the divide and conquer approach is illustrated by the fact that its kernel limit error is higher at high SNR, but lower at low SNR. This is of practical relevance, and is much related to the beneficial effect of *bagging* in noisy dataset scenarios. The divide and conquer approach, which only differs from bagging by the fact that the different partitions of the dataset are disjoint, was shown to reach bagging-like performance in various setups such as decision trees and neural networks [1].

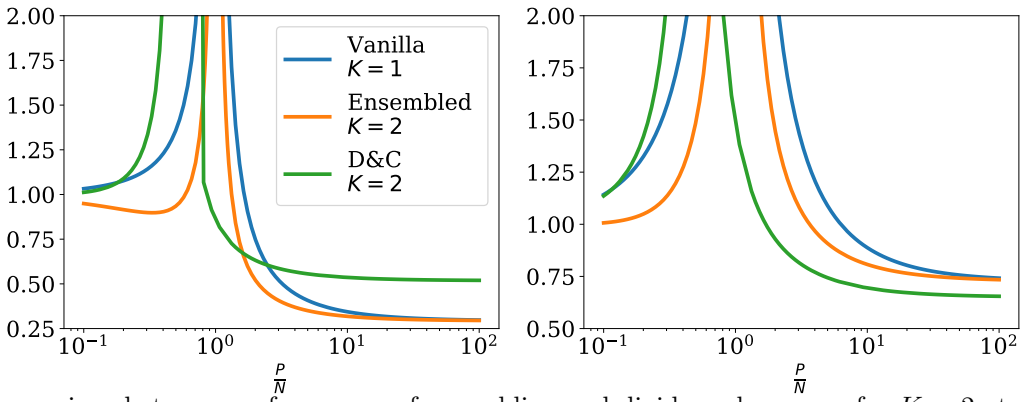

Figure 2: Comparison between performances of ensembling and divide and conquer for $K = 2$ at different SNR. **Left:** SNR=$\frac{F}{\tau} = 10$. **Right:** SNR=$\frac{F}{\tau} = 1$ Computations performed with fixed $N/D = 1$, $F = 1$ and $\lambda = 10^{-5}$.

1.3 Is it always better to be overparametrized ?

A common thought is that the double descent curve always reaches its minimum in the over-parametrized regime, leading to the idea that the corresponding model "cannot overfit". In this section, we show that this is not always the case. Three factors tend to shift the optimal generalization to the underparametrized regime: (i) increasing the numbers of learners from which we average the predictions, K, (ii) decreasing the signal-to-noise ratio (SNR), F/τ, and (iii) decreasing the size of the dataset, N/D. In other words, when ensembling on a small, noisy dataset, one is better off using an underparametrized model.
These three effects are shown in figure 3. In the left panel, we see that as we increase K, the minimum of test error jumps to the underparametrized regime $P < N$ for a high enough value of K. In the central/right panels, a similar effect occurs when decreasing the SNR or decreasing N/D.

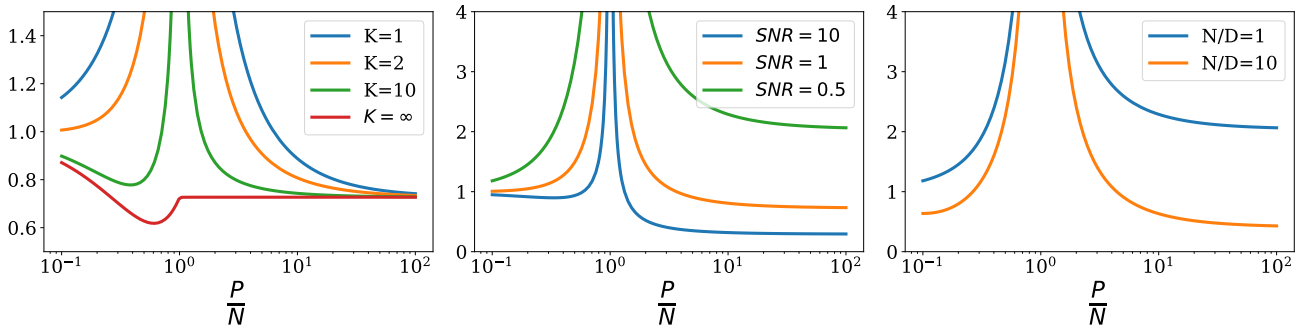

Figure 3: Generalisation error as a function of P/N: depending on the values of K, F/τ and N/D, optimal generalization can be reached in the underparametrized regime or the overparametrized regime. **Left:** $F/\tau = 1$, $N/D = 1$ and we vary K. This is the same as figure 5 in the main text, except that the higher noise causes the ensembling curve at $K \to \infty$ to exhibit a *dip* in the underparametrized regime. **Center:** $K = 2$, $N/D = 1$ and we vary F/τ. **Right:** $F/\tau = 1$, $K = 2$ and we vary N/D.

2 Statement of the Main Result

2.1 Assumptions

First, we state precisely the assumptions under which our main result is valid. Note, that these are the same as in [4].
Assumption 1: $\sigma : \mathbb{R} \to \mathbb{R}$ is a weakly differential function with derivative σ'. Assume there exists $c_0, c_1 < \infty \in \mathbb{R}$ such that for all $u \in \mathbb{R}$ $|\sigma(u)|, |\sigma'(u)| \leq c_0 \, \mathrm{e}^{c_1 |u|}$. Then define:

$$\mu_0 = \mathbb{E}\left[\sigma(u)\right] \quad \mu_1 = \mathbb{E}\left[u\sigma(u)\right] \quad \mu_\star^2 = \mathbb{E}\left[\sigma^2(u)\right] - \mu_0^2 - \mu_1^2, \tag{2}$$

where the expectation is over $u \sim \mathcal{N}(0,1)$. To facilitate readability, we specialize to the case $\mu_0 = 0$. This simply amounts to a shift ctivation function $\tilde{\sigma}$ of the network, $\tilde{\sigma}(x) = \sigma(x) - \mu_0$.

Assumption 2: We work in the high-dimensional limit, i.e. in the limit where the input dimension D, the hidden layer dimension P and the number of training points N go to infinity with their ratios fixed. That is:

$$N, P, D \to \infty, \qquad \frac{P}{D} \equiv \psi_1 = \mathcal{O}(1), \quad \frac{N}{D} \equiv \psi_2 = \mathcal{O}(1). \tag{3}$$

This condition implies that, in the computation of the risk $\mathcal{R}$, we can neglect all the terms of order $\mathcal{O}(1)$ in favour of the terms of order $\mathcal{O}(D)$.

Assumption 3: The labels are given by a linear ground truth, or teacher function:

$$y_\mu = f_d(\boldsymbol{X}_\mu) + \epsilon_\mu, \quad f_d(\boldsymbol{x}) = \langle \boldsymbol{\beta}, \boldsymbol{x} \rangle, \quad ||\boldsymbol{\beta}|| = F, \quad \epsilon_\mu \sim \mathcal{N}(0, \tau)\,. \tag{4}$$

Note that as explained in [4], it is easy to add a non linear component to the teacher, but the latter would not be captured by the model (the student) in the regime $N/D = \mathcal{O}(1)$, and would simply amount to an extra noise term.

2.2 Results

Here we give the explicit form of the quantities appearing in our main result. In these expressions, the index $a \in \{v, e, d\}$ distinguishes the *vanilla*, *ensembling* and *divide and conquer* terms.

$$\begin{aligned}
\Psi_1 &= \frac{1}{D} \operatorname{Tr}\left[H\left[S^v\right]^{-1} H\left[P_{\Psi_1}\right] \right], \\
\Psi_2^v &= \frac{1}{D} \operatorname{Tr}\left[H\left[S^v\right]^{-1} H\left[P_{\Psi_2^v}\right] \right], \\
\Psi_3^v &= \frac{1}{D} \operatorname{Tr}\left[H\left[S^v\right]^{-1} H\left[P_{\Psi_3^v}\right] \right], \\
\Psi_2^e &= \frac{1}{D} \operatorname{Tr}\left[H\left[S^e\right]^{-1} H\left[P_{\Psi_2^e}\right] \right], \\
\Psi_3^e &= \frac{1}{D} \operatorname{Tr}\left[H\left[S^e\right]^{-1} H\left[P_{\Psi_3^e}\right] \right], \\
\Psi_2^d &= \frac{1}{D} \operatorname{Tr}\left[H\left[S^d\right]^{-1} H\left[P_{\Psi_2^d}\right] \right],
\end{aligned}$$

where the Hessian matrix $H[F]$, for a given function $F : (q, r, \tilde{q}, \tilde{r}) \mapsto \mathbb{R}$ is defined as:

$$H\left[F\right] = \left[\begin{array}{cccc}
\frac{\partial F}{\partial q \partial q} & \frac{\partial F}{\partial q \partial r} & \frac{\partial F}{\partial q \partial \tilde{q}} & \frac{\partial F}{\partial q \partial \tilde{r}} \\
\frac{\partial F}{\partial q \partial r} & \frac{\partial F}{\partial r \partial r} & \frac{\partial F}{\partial r \partial \tilde{q}} & \frac{\partial F}{\partial r \partial \tilde{r}} \\
\frac{\partial F}{\partial q \partial \tilde{q}} & \frac{\partial F}{\partial r \partial \tilde{q}} & \frac{\partial F}{\partial \tilde{q} \partial \tilde{q}} & \frac{\partial F}{\partial \tilde{q} \partial \tilde{r}} \\
\frac{\partial F}{\partial q \partial \tilde{r}} & \frac{\partial F}{\partial r \partial \tilde{r}} & \frac{\partial F}{\partial \tilde{q} \partial \tilde{r}} & \frac{\partial F}{\partial \tilde{r} \partial \tilde{r}}
\end{array} \right]_{\substack{q=q^* \\ r=r^* \\ \tilde{r}=0 \\ \tilde{q}=0}},$$

with $q^\star$ and $r^\star$ being the solutions of the fixed point equation for the function $S_0 : (q, r) \mapsto \mathbb{R}$ defined below:

$$\begin{cases} \frac{\partial S_0(q,r)}{\partial q} = 0 \\ \frac{\partial S_0(q,r)}{\partial r} = 0. \end{cases}$$

$$S_0(q, r) = \lambda \psi_1^2 \psi_2 q + \psi_2 \log\left(\frac{\mu_\star^2 \psi_1 q}{\mu_1^2 \psi_1 r + 1} + 1 \right) + \frac{r}{q} + (1 - \psi_1) \log(q) + \psi_2 \log\left(\mu_1^2 \psi_1 r + 1\right) - \log(r).$$

The explicit expression of the above quantities in terms of $(q, r, \tilde{q}, \tilde{r})$ is given below.

2.3 Explicit expression of S^v, S^e, S^d

Here we present the explicit formulas for S^v, S^e, S^d, which are defined as the functions $(q, r, \tilde{q}, \tilde{r}) \mapsto \mathbb{R}$ such that:

$$
\begin{aligned}
S^v(q,r,\tilde{q},\tilde{r}) &= 2\left(S_0(q,r) + \tilde{q} f^v(q,r) + \tilde{r} g^v(q,r)\right) \\
S^e(q,r,\tilde{q},\tilde{r}) &= S_0(q,r) + \tilde{r}^2 f^e(q,r) + \tilde{q}^2 g^e(q,r) \\
S^d(q,r,\tilde{q},\tilde{r}) &= S_0(q,r) + \tilde{r}^2 f^d(q,r) + \tilde{q}^2 g^d(q,r),
\end{aligned}
$$

where we defined the functions $(q, r) \mapsto \mathbb{R}$,

$$
\begin{aligned}
f^v(q,r) &= \lambda \psi_1^2 \psi_2 + \frac{\mu_\star^2 \psi_1 \psi_2}{\mu_\star^2 \psi_1 q + \mu_1^2 \psi_1 r + 1} + \frac{1-\psi_1}{q} - \frac{r}{q^2}, \\
g^v(q,r) &= -\frac{\mu_\star^2 \mu_1^2 \psi_1^2 \psi_2 q}{\left(\mu_1^2 \psi_1 r + 1\right)\left(\mu_\star^2 \psi_1 q + \mu_1^2 \psi_1 r + 1\right)} + \frac{\mu_1^2 \psi_1 \psi_2}{\mu_1^2 \psi_1 r + 1} + \frac{1}{q} - \frac{1}{r}, \\
f^e(q,r) &= \frac{2 r \mu_1^2 \psi_1 \left(1 + q\mu_\star^2 \psi_1\right) + \left(1 + q\mu_\star^2 \psi_1\right)^2 - r^2 \mu_1^4 \psi_1^2 (-1 + \psi_2)}{r^2 \left(1 + r\mu_1^2 \psi_1 + q\mu_\star^2 \psi_1\right)^2}, \\
g^e(q,r) &= \frac{\psi_1}{q^2}, \\
f^d(q,r) &= \frac{1}{r^2}, \\
g^d(q,r) &= \frac{\psi_1}{q^2}.
\end{aligned}
$$

2.4 Explicit expression of $P_{\Psi_1}, P^v_{\Psi_2}, P^v_{\Psi_3}, P^e_{\Psi_2}, P^e_{\Psi_3}, P^d_{\Psi_2}$

Here we present the explicit formulas for $P_{\Psi_1}, P^v_{\Psi_2}, P^v_{\Psi_3}, P^e_{\Psi_2}, P^e_{\Psi_3}, P^d_{\Psi_2}$, which are defined as the functions $(q, r, \tilde{q}, \tilde{r}) \mapsto \mathbb{R}$ such that:

$$
\begin{aligned}
P_{\Psi_1} &= \psi_1 \psi_2 \mu_1^2 \left(M_X^{11} + \mu_1^2 \mu_\star^2 \psi_1^2 \left(M_X M_W^v M_X\right)^{11} + \mu_\star^2 \psi_1 \left(M_X M_W^v\right)^{11}\right), \\
P_{\Psi_2^v} &= D\psi_1^2 \psi_2 (\mu_1^2 \tilde{r} + \mu_\star^2 \tilde{q}) \left[\mu_1^2 P^v_{XX} - 2\mu_1^2 \mu_\star^2 \psi_1 P^v_{WX} + \mu_\star^2 P^v_{WW}\right], \\
P_{\Psi_3^v} &= D\psi_1^2 \psi_2 (\mu_1^2 \tilde{r} + \mu_\star^2 \tilde{q}) \left[\mu_1^2 \left(M_X^{12} + \mu_1^2 \mu_\star^2 \psi_1^2 \left[M_X M_W^v M_X\right]^{12}\right) - 2\mu_1^2 \mu_\star^2 \psi_1 \left[M_X M_W^v\right]^{12} + \mu_\star^2 \left[M_W^v\right]^{12}\right], \\
P_{\Psi_2^e} &= D\psi_1^2 \psi_2 \mu_1^2 \tilde{r} \left[\mu_1^2 P^e_{XX} - 2\mu_1^2 \mu_\star^2 \psi_1 P^e_{WX} + \mu_\star^2 P^e_{WW}\right], \\
P_{\Psi_3^e} &= D\psi_1^2 \psi_2 \mu_1^2 \tilde{r} \left[\mu_1^2 \left(M_X^{12} + \mu_1^2 \mu_\star^2 \psi_1^2 \left[M_X M_W^e M_X\right]^{12}\right) - 2\mu_1^2 \mu_\star^2 \psi_1 \left[M_X M_W^e\right]^{12} + \mu_\star^2 [M_W^e]^{12}\right], \\
P_{\Psi_2^d} &= D\mu_1^2 \psi_1 \psi_2^2 \tilde{r} \left[\psi_1 \mu_1^2 P_{XX} + 2\mu_\star^2 \mu_1^2 \psi_1^2 P_{WX} + \mu_\star^2 \psi_1 P_{WW}\right],
\end{aligned}
$$

where we defined the scalars P_{XX}, P_{WX}, P_{WW} as follows:

$$
\begin{aligned}
P^v_{XX} &= \psi_2 N_X^{12} + M_X^{12} + 2\psi_2 (\mu_1 \mu_\star \psi_1)^2 \left[M_X N_X M_W^a\right]^{12} + (\mu_1 \mu_\star \psi_1)^2 \left[M_X M_W^a M_X\right]^{12} \\
&\quad + \psi_2 (\mu_1 \mu_\star \psi_1)^4 \left[M_X M_W N_X M_W^a M_X\right]^{12}, \\
P^v_{WX} &= \psi_2 \left[N_X M_W^a\right]^{12} + \left[M_X M_W^a\right]^{12} + \psi_2 (\mu_1 \mu_\star \psi_1)^2 \left[M_X M_W^a N_X M_W^a\right]^{12}, \\
P^v_{WW} &= \left[M_W^a\right]^{12} + \psi_2 (\mu_1 \mu_\star \psi_1)^2 \left[M_W^a N_X M_W^a\right]^{12}, \\
P^e_{XX} &= P^v_{XX}, \\
P^e_{WX} &= P^v_{WX}, \\
P^e_{WW} &= P^v_{WW}, \\
P^d_{XX} &= \left(N_X^{d11} + 2(\mu_1 \mu_\star \psi_1)^2 \left[N_X^d M_W^d M_X^d\right]^{11} + (\mu_1 \mu_\star \psi_1)^4 \left[M_X^d M_W^d N_X^d M_W^d M_X^d\right]^{11}\right), \\
P^d_{WX} &= \left[N_X^d M_W^d\right]^{11} + (\mu_1 \mu_\star \psi_1)^2 \left[M_X^d M_W^d N_X^d M_W^d\right]^{11}, \\
P^d_{WW} &= (\mu_1 \mu_\star \psi_1)^2 \left[M_W^d N_X^d M_W^d\right]^{11},
\end{aligned}
$$

and the 2×2 matrices M_X, M_W, N_X as follows:

$$M_X^v = \begin{bmatrix} \frac{r}{1+\mu_1^2\psi_1 r} & \frac{\tilde{r}}{\left(1+\mu_1^2\psi_1 r\right)^2} \\ \frac{\tilde{r}}{\left(1+\mu_1^2\psi_1 r\right)^2} & \frac{r}{1+\mu_1^2\psi_1 r} \end{bmatrix}, \quad M_W^v = \begin{bmatrix} \frac{q\left(1+\mu_1^2\psi_1 r\right)}{1+2\mu_1^2\psi_1 r+\mu_\star^2\psi_1 q} & \frac{q^2\mu_1^2\mu_\star^2\psi_1^2\tilde{r}}{\left(1+\mu_1^2\psi_1 r+\mu_\star^2\psi_1 q\right)^2} \\ \frac{q^2\mu_1^2\mu_\star^2\psi_1^2\tilde{r}}{\left(1+\mu_1^2\psi_1 r+\mu_\star^2\psi_1 q\right)^2} & \frac{q\left(1+\mu_1^2\psi_1 r\right)}{1+2\mu_1^2\psi_1 r+\mu_\star^2\psi_1 q} \end{bmatrix}, \quad N_X^v = \frac{1}{\left(1+\mu_1^2\psi_1 r\right)^2}\begin{bmatrix} r & \tilde{r} \\ \tilde{r} & r \end{bmatrix},$$

$$M_X^e = M_X^v, \quad M_W^e = M_W^v + \frac{(1+r\mu_1^2\psi_1)^2\tilde{q}}{\left(1+\mu_1^2\psi_1 r+\mu_\star^2\psi_1 q\right)^2}\begin{bmatrix} 0 & 1 \\ 1 & 0 \end{bmatrix}, \quad N_X^e = \frac{1}{\left(1+\mu_1^2\psi_1 r\right)^2}\begin{bmatrix} r & \tilde{r} \\ \tilde{r} & r \end{bmatrix},$$

$$M_X^d = \frac{r}{1+\mu_1^2\psi_1 r^2}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad M_W^d = \frac{q(1+\mu_1^2\psi_1 r)}{1+\mu_1^2\psi_1 r+\mu_\star^2\psi_1 q}\begin{bmatrix} 1 & 0 \\ 0 & 1 \end{bmatrix}, \quad N_X^d = \frac{1}{(1+\mu_1^2\psi_1 r)^2}\begin{bmatrix} \tilde{r} & 0 \\ 0 & \tilde{r} \end{bmatrix}.$$

3 Replica Computation

3.1 Toolkit

3.1.1 Gaussian integrals

In order to obtain the main result for the generalisation error, we perform the averages over all the sources of randomness in the system in the following order: over the dataset X, then over the noise W, and finally over the random feature layers Θ. Here are some useful formulaes used throughout the computations:

$$\begin{cases} \int e^{-\frac{1}{2}x_i G_{ij} x_j + J_i x_i} dx & = (\det G)^{-\frac{1}{2}} e^{\frac{1}{2}J_i G_{ij}^{-1} J_j}, \\ \int x_a e^{-\frac{1}{2}x_i G_{ij} x_j + J_i x_i} dx & = P_a^1 (\det G)^{-\frac{1}{2}} e^{\frac{1}{2}J_i G_{ij}^{-1} J_j}, \\ \int x_a x_b e^{-\frac{1}{2}x_i G_{ij} x_j + J_i x_i} dx & = P_{ab}^2 (\det G)^{-\frac{1}{2}} e^{\frac{1}{2}J_i G_{ij}^{-1} J_j}, \\ \int x_a x_b x_c e^{-\frac{1}{2}x_i G_{ij} x_j + J_i x_i} dx & = P_{abc}^3 (\det G)^{-\frac{1}{2}} e^{\frac{1}{2}J_i G_{ij}^{-1} J_j}, \\ \int x_a x_b x_c x_d e^{-\frac{1}{2}x_i G_{ij} x_j + J_i x_i} dx & = P_{abcd}^4 (\det G)^{-\frac{1}{2}} e^{\frac{1}{2}J_i G_{ij}^{-1} J_j}, \end{cases} \tag{5}$$

with

$$\begin{aligned} P_a^1 &= [G^{-1}J]_a, \\ P_{ab}^2 &= ((G^{-1})_{ab} + [G^{-1}J]_a [G^{-1}J]_b), \\ P_{abc}^3 &= \sum_{\overline{a},\overline{b},\overline{c} \in perm(abc)} \left((G^{-1})_{\overline{a}\overline{b}} [G^{-1}J]_{\overline{c}} + [G^{-1}J]_{\overline{a}} [G^{-1}J]_{\overline{b}} [G^{-1}J]_{\overline{c}}\right), \\ P_{abcd}^4 &= \sum_{\overline{a},\overline{b},\overline{c},\overline{d} \in perm(abcd)} \left((G^{-1})_{\overline{a}\overline{b}} (G^{-1})_{\overline{c}\overline{d}} + [G^{-1}J]_{\overline{a}} [G^{-1}J]_{\overline{b}} [G^{-1}J]_{\overline{c}} [G^{-1}J]_{\overline{d}} + (G^{-1})_{\overline{a}\overline{b}} [G^{-1}J]_{\overline{c}} [G^{-1}J]_{\overline{d})}\right). \end{aligned}$$

3.1.2 Replica representation of an inverse matrix

To obtain gaussian integrals we will use the "replica" representation the element (ij) of a matrix M of size D:

$$M_{ij}^{-1} = \lim_{n \to 0} \int \left(\prod_{\alpha=1}^{n} \prod_{i=1}^{D} d\eta_i^\alpha \right) \eta_i^1 \eta_j^1 \exp\left(-\frac{1}{2} \eta_i^\alpha M_{ij} \eta_j^\alpha \right). \tag{6}$$

Indeed, using the gaussian integral representation of the inverse of M,

$$\begin{aligned} M_{ij}^{-1} &= \mathcal{Z}^{-1} \int \left(\prod_{i=1}^{D} d\eta_i \right) \eta_i \eta_j \exp\left(-\frac{1}{2} \eta_i M_{ij} \eta_j \right), \\ \mathcal{Z} &= \sqrt{\frac{(2\pi)^D}{\det M}} \\ &= \int \left(\prod_{i=1}^{D} d\eta_i \right) \exp\left(-\frac{1}{2} \eta_i M_{ij} \eta_j \right). \end{aligned}$$

Using the replica identity, we rewrite this as

$$M_{ij}^{-1} = \lim_{n\to 0} \mathcal{Z}^{n-1} \int \left(\prod_{i=1}^{D} d\eta_i\right) \eta_i \eta_j \exp\left(-\frac{1}{2}\eta_i M_{ij} \eta_j\right).$$

Renaming the integration variable of the integral on the left as η^1 and the $n-1$ others as $\eta^\alpha, \alpha \in \{2, n\}$, we obtain expression (6).

3.2 The Random Feature model

In what follows, we will explicitly leave the indices of all the quantities used. We use the notation, called Einstein summation convention in physics, in which all repeated indices are summed but the sum is not explicitly written. Indices $i \in \{1...D\}$ are used to refer to the input dimension, $h \in \{1...P\}$ to refer to the hidden layer dimension and $\mu \in \{1...N\}$ to refer to the number of data points.

3.2.1 With a single learner

In the random features model, the predictor can be computed explicitly:

$$\hat{a} = \frac{1}{\sqrt{D}} \boldsymbol{y}_\mu^{\mathrm{T}} \boldsymbol{Z}_{\mu h} \left(\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{Z} + \psi_1 \psi_2 \lambda \mathbf{I}_N\right)^{-1}_{hh'} \tag{7}$$

$$f(\boldsymbol{x}) = \hat{a}_h \sigma\left(\frac{\boldsymbol{\Theta}_{h'i} \boldsymbol{x}_i}{\sqrt{D}}\right) \tag{8}$$

$$= \boldsymbol{y}_\mu^{\mathrm{T}} \boldsymbol{Z}_{\mu h} \left(\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{Z} + \psi_1 \psi_2 \lambda \mathbf{I}_N\right)^{-1}_{hh'} \sigma\left(\frac{\boldsymbol{\Theta}_{h'i} \boldsymbol{x}_i}{\sqrt{D}}\right) / \sqrt{D}, \tag{9}$$

where

$$\boldsymbol{y}_\mu = f_d\left(\boldsymbol{X}_\mu\right) + \boldsymbol{\epsilon}_\mu, \tag{10}$$

$$\boldsymbol{Z}_{\mu h} = \frac{1}{\sqrt{D}} \sigma\left(\frac{1}{\sqrt{D}} \boldsymbol{\Theta}_{hi} \boldsymbol{X}_{\mu i}\right). \tag{11}$$

Hence the test error can be computed as:

$$\mathcal{R}_{\mathrm{RF}} = \mathbb{E}_{\boldsymbol{x}} \left[\left(f_d(\boldsymbol{x}) - \boldsymbol{y}_\mu^{\mathrm{T}} \boldsymbol{Z}_{\mu h} \left(\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{Z} + \psi_1 \psi_2 \lambda \mathbf{I}_N\right)^{-1}_{hh'} \sigma\left(\frac{\boldsymbol{\Theta}_{h'i} \boldsymbol{x}_i}{\sqrt{D}}\right) / \sqrt{D}\right)^2\right] \tag{12}$$

$$= \mathbb{E}_{\boldsymbol{x}} \left[f_d(\boldsymbol{x})^2\right] - 2 \boldsymbol{y}_\mu^{\top} \boldsymbol{Z}_{\mu h} \left(\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{Z} + \psi_1 \psi_2 \lambda \mathbf{I}_N\right)^{-1}_{hh'} \boldsymbol{V}_{h'} / \sqrt{D} + \boldsymbol{y}_\mu^{\mathrm{T}} \boldsymbol{Z}_{\mu h} \left(\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{Z} + \psi_1 \psi_2 \lambda \mathbf{I}_N\right)^{-1}_{hh_1} \boldsymbol{U}_{h_1 h_2} \left(\boldsymbol{Z}^{\mathrm{T}} \boldsymbol{Z} + \psi_1 \psi_2 \lambda \mathbf{I}_N\right)^{-1}_{h_2 h'} \boldsymbol{Z}^{\mathrm{T}}_{h'\mu'} \boldsymbol{y}_{\mu'} / D, \tag{13}$$

where

$$\boldsymbol{V}_h = \mathbb{E}_{\boldsymbol{x}} \left[f_d(\boldsymbol{x}) \sigma\left(\frac{\langle \boldsymbol{\Theta}_{hi} \boldsymbol{x}_i \rangle}{\sqrt{D}}\right)\right], \tag{14}$$

$$\boldsymbol{U}_{hh'} = \mathbb{E}_{\boldsymbol{x}} \left[\sigma\left(\frac{\langle \boldsymbol{\Theta}_{hi} \boldsymbol{x}_i \rangle}{\sqrt{D}}\right) \sigma\left(\frac{\boldsymbol{\Theta}_{h'i}, \boldsymbol{x}_i}{\sqrt{D}}\right)\right]. \tag{15}$$

3.2.2 Ensembling over K learners

When ensembling over K learners with independently sampled random feature vectors, the predictor becomes:

$$f(\boldsymbol{x}) = \frac{1}{K\sqrt{D}} \sum_k \boldsymbol{y}_\mu^{\mathrm{T}} \boldsymbol{Z}^{(k)}_{\mu h} \left(\boldsymbol{Z}^{\mathrm{T}(k)} \boldsymbol{Z}^{(k)} + \psi_1 \psi_2 \lambda \mathbf{I}_N\right)^{-1}_{hh'} \sigma\left(\frac{\boldsymbol{\Theta}^{(k)}_{h'i} \boldsymbol{x}_i}{\sqrt{D}}\right), \tag{16}$$

where

$$\boldsymbol{Z}^{(k)}_{\mu h} = \frac{1}{\sqrt{D}} \sigma\left(\frac{1}{\sqrt{D}} \boldsymbol{\Theta}^{(k)}_{hi} \boldsymbol{X}_{\mu i}\right). \tag{17}$$

The generalisation error is then given by:

$$\mathcal{R}_{\rm RF} = \mathbb{E}_{\boldsymbol{x}}\left[\left(f_d(\boldsymbol{x}) - \frac{1}{K}\sum_k \boldsymbol{y}^{\rm T}\boldsymbol{Z}^{(k)}\left(\boldsymbol{Z}^{{\rm T}(k)}\boldsymbol{Z}^{(k)} + \psi_1\psi_2\lambda\mathbf{I}_N\right)^{-1}\sigma\left(\frac{\boldsymbol{\Theta}^{(k)}_{h'i}\boldsymbol{x}_i}{\sqrt{D}}\right)/\sqrt{D}\right)^2\right] \tag{18}$$

$$= \mathbb{E}_{\boldsymbol{x}}\left[f_d(\boldsymbol{x})^2\right] - \frac{2}{K}\sum_k \boldsymbol{y}^{\top}\boldsymbol{Z}^{(k)}\left(\boldsymbol{Z}^{{\rm T}(k)}\boldsymbol{Z}^{(k)} + \psi_1\psi_2\lambda\mathbf{I}_N\right)^{-1}\boldsymbol{V}^{(k)}/\sqrt{D}$$

$$+ \frac{1}{K^2}\sum_{\substack{k\\ l\neq k}} \boldsymbol{y}^{\rm T}\boldsymbol{Z}^{(k)}\left(\boldsymbol{Z}^{{\rm T}(k)}\boldsymbol{Z}^{(k)} + \psi_1\psi_2\lambda\mathbf{I}_N\right)^{-1}\boldsymbol{U}^{(kl)}\left(\boldsymbol{Z}^{{\rm T}(l)}\boldsymbol{Z}^{(l)} + \psi_1\psi_2\lambda\mathbf{I}_N\right)^{-1}\boldsymbol{Z}^{(\boldsymbol{l})}\mathbf{T}\boldsymbol{y}/D, \tag{19}$$

where

$$\boldsymbol{V}^{(k)}_h = \mathbb{E}_{\boldsymbol{x}}\left[f_d(\boldsymbol{x})\sigma\left(\frac{\langle\boldsymbol{\Theta}^{(k)}_{hi}\boldsymbol{x}_i\rangle}{\sqrt{D}}\right)\right], \tag{20}$$

$$\boldsymbol{U}^{(kl)}_{hh'} = \mathbb{E}_{\boldsymbol{x}}\left[\sigma\left(\frac{\langle\boldsymbol{\Theta}^{(k)}_{hi}\boldsymbol{x}_i\rangle}{\sqrt{D}}\right)\sigma\left(\frac{\langle\boldsymbol{\Theta}^{(l)}_{h'i'}\boldsymbol{x}_{i'}\rangle}{\sqrt{D}}\right)\right]. \tag{21}$$

3.2.3 Equivalent Gaussian Covariate Model

It was shown in [4] that the random features model is equivalent, in the high-dimensional limit of Assumption 2, to a Gaussian covariate model in which the activation function σ is replaced as:

$$\sigma\left(\frac{\boldsymbol{\Theta}^{(k)}_{hi}\boldsymbol{X}_{\mu i}}{\sqrt{D}}\right) \to \mu_0 + \mu_1\frac{\boldsymbol{\Theta}^{(k)}_{hi}\boldsymbol{X}_{\mu i}}{\sqrt{D}} + \mu_\star\boldsymbol{W}^{(k)}_{\mu h}, \tag{22}$$

with $\boldsymbol{W}^{(k)} \in \mathbb{R}^{N\times P}$, $W^{(k)}_{\mu h} \sim \mathcal{N}(0,1)$ and μ_0, μ_1 and $\mu_\star$ defined in (2). To simplify the calculations, we take $\mu_0 = 0$, which amounts to adding a constant term to the activation function σ.
This powerful mapping allows to express the quantities $\boldsymbol{U}, \boldsymbol{V}$. We will not repeat their calculations here: the only difference here is $\boldsymbol{U}^{kl}$, which carries extra indices k, l due to the different initialization of the random features $\boldsymbol{\Theta}^{(k)}$. In our case,

$$\boldsymbol{U}^{(kl)}_{hh'} = \frac{\mu_1^2}{D}\boldsymbol{\Theta}^{(k)}_{hi}\boldsymbol{\Theta}^{(l)}_{h'i} + \mu_\star^2\delta_{kl}\delta_{hh'}. \tag{23}$$

Hence we can rewrite the test error as

$$\mathop{\mathbb{E}}_{\{\boldsymbol{\Theta}^{(k)}\},\boldsymbol{X},\boldsymbol{\varepsilon}}\left[\mathcal{R}_{\rm RF}\right] = F^2\left(1 - 2\Psi_1^v\right) + \frac{1}{K}\left(F^2\Psi_2^v + \tau^2\Psi_3^v\right) + \left(1 - \frac{1}{K}\right)\left(F^2\Psi_2^e + \tau^2\Psi_3^e\right), \tag{24}$$

where Ψ_1,Ψ_2^v,Ψ_2^e,Ψ_3^v,Ψ_3^e are given by:

$$\Psi_1 = \frac{1}{D}\,{\rm Tr}\left[\left(\frac{\mu_1}{D}\boldsymbol{X}\boldsymbol{\Theta}^{(1)\top}\right)^{\top}\boldsymbol{Z}^{(\boldsymbol{1})}\left(\boldsymbol{Z}^{(\boldsymbol{1})\top}\boldsymbol{Z}^{(\boldsymbol{1})} + \psi_1\psi_2\lambda\mathbf{I}_N\right)^{-1}\right],$$

$$\Psi_2^v = \frac{1}{D}\,{\rm Tr}\left[\left(\boldsymbol{Z}^{(\boldsymbol{1})\top}\boldsymbol{Z}^{(\boldsymbol{1})} + \psi_1\psi_2\lambda\mathbf{I}_N\right)^{-1}\left(\frac{\mu_1^2}{D}\boldsymbol{\Theta}^{(1)}\boldsymbol{\Theta}^{(1)\top} + \mu_*^2\mathbf{I}_N\right)\left(\boldsymbol{Z}^{(\boldsymbol{1})\top}\boldsymbol{Z}^{(\boldsymbol{1})} + \psi_1\psi_2\lambda\mathbf{I}_N\right)^{-1}\boldsymbol{Z}^{(\boldsymbol{1})\top}\left(\frac{1}{D}\boldsymbol{X}\boldsymbol{X}^{\top}\right)\boldsymbol{Z}^{(\boldsymbol{1})}\right],$$

$$\Psi_3^v = \frac{1}{D}\,{\rm Tr}\left[\left(\boldsymbol{Z}^{(\boldsymbol{1})\top}\boldsymbol{Z}^{(\boldsymbol{1})} + \psi_1\psi_2\lambda\mathbf{I}_N\right)^{-1}\left(\frac{\mu_1^2}{D}\boldsymbol{\Theta}^{(1)}\boldsymbol{\Theta}^{(1)\top} + \mu_\star^2\mathbf{I}_N\right)\left(\boldsymbol{Z}^{(\boldsymbol{1})\top}\boldsymbol{Z}^{(\boldsymbol{1})} + \psi_1\psi_2\lambda\mathbf{I}_N\right)^{-1}\boldsymbol{Z}^{(\boldsymbol{1})\top}\boldsymbol{Z}^{(\boldsymbol{1})}\right],$$

$$\Psi_2^e = \frac{1}{D}\,{\rm Tr}\left[\left(\boldsymbol{Z}^{(\boldsymbol{1})\top}\boldsymbol{Z}^{(\boldsymbol{1})} + \psi_1\psi_2\lambda\mathbf{I}_N\right)^{-1}\left(\frac{\mu_1^2}{D}\boldsymbol{\Theta}^{(1)}\boldsymbol{\Theta}^{(2)\top}\right)\left(\boldsymbol{Z}^{(\boldsymbol{2})\top}\boldsymbol{Z}^{(\boldsymbol{2})} + \psi_1\psi_2\lambda\mathbf{I}_N\right)^{-1}\boldsymbol{Z}^{(\boldsymbol{2})\top}\left(\frac{1}{D}\boldsymbol{X}\boldsymbol{X}^{\top}\right)\boldsymbol{Z}^{(\boldsymbol{1})}\right],$$

$$\Psi_3^e = \frac{1}{D}\,{\rm Tr}\left[\left(\boldsymbol{Z}^{(\boldsymbol{1})\top}\boldsymbol{Z}^{(\boldsymbol{1})} + \psi_1\psi_2\lambda\mathbf{I}_N\right)^{-1}\left(\frac{\mu_1^2}{D}\boldsymbol{\Theta}^{(1)}\boldsymbol{\Theta}^{(2)\top}\right)\left(\boldsymbol{Z}^{(\boldsymbol{2})\top}\boldsymbol{Z}^{(\boldsymbol{2})} + \psi_1\psi_2\lambda\mathbf{I}_N\right)^{-1}\boldsymbol{Z}^{(\boldsymbol{2})\top}\boldsymbol{Z}^{(\boldsymbol{1})}\right].$$

3.3 Computation of the *vanilla* terms

To start with, let us compute the vanilla terms (those who carry a superscript v), which involve a single instance of the random feature vectors. Note that these were calculated in [4] by evaluating the Stieljes transform of the

random matrices of which we need to calculate the trace. The replica method used here makes the calculation of the vanilla terms carry over easily to the the ensembling terms (superscript e) and the divide and conquer term (superscript d). To illustrate the calculation steps, we will calculate Ψ_3^v, then provide the results for Ψ_2^v and Ψ_1. In the vanilla terms, the two inverse matrices that appear are the same. Hence we use twice the replica identity (6), introducing $2n$ replicas which all play the same role:

$$M_{ij}^{-1}M_{kl}^{-1} = \lim_{n\to 0}\int\left(\prod_{\alpha=1}^{2n} d\eta\right)\eta_i^1\eta_j^1\eta_k^2\eta_l^2\exp\left(-\frac{1}{2}\eta^\alpha M_{ij}\eta^\alpha\right). \tag{25}$$

The first step is to perform the averages, i.e. the Gaussian integrals, over the dataset X, the deterministic noise W induced by the non-linearity of the activation function and the random features Θ.

3.3.1 Averaging over the dataset

Replacing the activation function by its Gaussian covariate equivalent model and using (25), the term Ψ_3 can be expanded as:

$$\begin{aligned}
\Psi_3^v &= \frac{1}{D}\left[\boldsymbol{Z}_{\mu h}\left(\boldsymbol{Z}^\top\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N\right)^{-1}_{hh_1}\left(\frac{\mu_1^2}{D}\boldsymbol{\Theta}_{h_1 i}\boldsymbol{\Theta}_{h_2 i}+\mu_\star^2\delta_{h_1h_2}\right)\left(\boldsymbol{Z}^\top\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N\right)^{-1}_{h_2h'}\boldsymbol{Z}_{h'\mu}\right]\\
&= \frac{1}{D}\left(\frac{\mu_1^2}{D}\boldsymbol{\Theta}_{h_1 i}\boldsymbol{\Theta}_{h_2 i}+\mu_\star^2\delta_{h_1h_2}\right)\left[\boldsymbol{Z}_{\mu h}\boldsymbol{Z}_{h'\mu}\right]\int\left(\prod_{\alpha=1}^{2n}d\eta^\alpha\right)\eta_h^1\eta_{h_1}^1\eta_{h_2}^2\eta_{h'}^2\exp\left(-\frac{1}{2}\eta_h^\alpha\left(\boldsymbol{Z}^\top\boldsymbol{Z}+\psi_1\psi_2\lambda\mathbf{I}_N\right)_{hh'}\eta_{h'}^\alpha\right)\\
&= \frac{1}{D^2}\left(\frac{\mu_1^2}{D}\boldsymbol{\Theta}_{h_1 i}\boldsymbol{\Theta}_{h_2 i}+\mu_\star^2\delta_{h_1h_2}\right)\left(\frac{\mu_1}{\sqrt{D}}\boldsymbol{\Theta}\boldsymbol{X}+\mu_\star\boldsymbol{W}\right)_{h\mu}\left(\frac{\mu_1}{\sqrt{D}}\boldsymbol{\Theta}\boldsymbol{X}+\mu_\star\boldsymbol{W}\right)_{h'\mu}\\
&\int\left(\prod_{\alpha=1}^{2n}d\eta^\alpha\right)\eta_h^1\eta_{h_1}^1\eta_{h_2}^2\eta_{h'}^2\exp\left(-\frac{1}{2}\eta_h^\alpha\left(\frac{1}{D}\left(\frac{\mu_1}{\sqrt{D}}\boldsymbol{\Theta}\boldsymbol{X}+\mu_\star\boldsymbol{W}\right)_{h\mu}\left(\frac{\mu_1}{\sqrt{D}}\boldsymbol{\Theta}\boldsymbol{X}+\mu_\star\boldsymbol{W}\right)_{h'\mu}+\psi_1\psi_2\lambda\delta_{hh'}\right)\eta_{h'}^\alpha\right).
\end{aligned}$$

Now, we introduce $\lambda_i^\alpha := \frac{1}{\sqrt{P}}\eta_h^\alpha\boldsymbol{\Theta}_{hi}$, and enforce this relation using the Fourier representation of the delta-function:

$$1 = \int d\lambda_i^\alpha d\hat{\lambda}_i^\alpha e^{i\hat{\lambda}_i^\alpha(\sqrt{P}\lambda_i^\alpha-\eta_h^\alpha\boldsymbol{\Theta}_{hi})}. \tag{26}$$

The average over the dataset $\boldsymbol{X}_{\mu i}$ has the form of (5) with:

$$(G_X)_{\mu\mu',ii'} = \delta_{\mu\mu'}\left(\delta_{ii'}+\frac{\mu_1^2\psi_1}{D}\lambda_i^\alpha\lambda_{i'}^\alpha\right), \tag{27}$$

$$(J_X)_{\mu,i} = \frac{\mu_1\mu_\star\sqrt{\psi_1}}{D}\sum_{\alpha a}\lambda_i^\alpha\eta_h^\alpha\boldsymbol{W}_{\mu h}. \tag{28}$$

Using formulae (5), we obtain:

$$\begin{aligned}
\Psi_3^v =& \frac{N}{D^2}\int\left(\prod_{\alpha=1}^{2n}d\eta^\alpha d\lambda^\alpha d\hat{\lambda}^\alpha\right)\left(\frac{\mu_1^2P}{D}\lambda_i^1\lambda_i^2+\mu_\star^2\eta_i^1\eta_i^2\right)\\
&\left[\mu_\star^2\eta_h^1\boldsymbol{W}_h\boldsymbol{W}_{h'}\eta_{h'}^2+\mu_1^2\psi_1\lambda_i^1\lambda_{i'}^2\left((G_X^{-1})_{ii'}+(G_X^{-1}J_X)_i(G_X^{-1}J_X)_{i'}\right)+2\mu_1\mu_\star\sqrt{\psi_1}\lambda_i^1\boldsymbol{W}_h\eta_h^2(G_X^{-1}J_X)_i\right]\\
&\exp\left(-\frac{n}{2}\log\det(G_X)-\frac{1}{2}\eta_h^\alpha\left(\frac{\mu_\star^2}{D}\boldsymbol{W}_h\boldsymbol{W}_{h'}\delta_{\alpha\beta}-\frac{\mu_1^2\mu_\star^2\psi_1}{D^2}\boldsymbol{W}_h\lambda_i^\alpha(G_X)^{-1}_{ii'}\lambda_{i'}^\beta\boldsymbol{W}_{h'}+\psi_1\psi_2\lambda\delta_{hh'}\right)\eta_{h'}^\beta+i\hat{\lambda}_i^\alpha(\sqrt{P}\lambda_i^\alpha-\eta_h^\alpha\boldsymbol{\Theta}_{hi})\right)
\end{aligned}$$

Note that due to with a slight abuse of notation we got rid of indices μ, which all sum up trivially to give a global factor N.

3.3.2 Averaging over the deterministic noise

The expectation over the deterministic noise $\boldsymbol{W}_h$ is a Gausssian integral of the form (5) with:

$$[G_W]_{hh'} = \delta_{hh'} + \frac{\mu_\star^2}{D}\eta_h^\alpha A^{\alpha\beta}\eta_{h'}^\beta, \tag{29}$$

$$[J_W]_h = 0, \tag{30}$$

$$A^{\alpha\beta} = \delta_{\alpha\beta} - \mu_1^2\psi_1\frac{1}{D}\sum_{i,j}\lambda_i^\alpha[G_X^{-1}]_{ij}\lambda_j^\beta. \tag{31}$$

Note that the prefactor involves, constant, linear and quadratic terms in $\boldsymbol{W}$ since:

$$(G_X^{-1}J_X)_i = \frac{\mu_1\mu_\star\sqrt{\psi_1}}{D}\left[\eta^\alpha \boldsymbol{W}\right]\left[G_X^{-1}\lambda^\alpha\right]_i.$$

Thus, one obtains:

$$\Psi_3^v = \frac{\psi_2}{D}\int\left(\prod_{\alpha=1}^{2n} d\eta^\alpha d\lambda^\alpha d\hat{\lambda}^\alpha\right)\left(\mu_1^2\psi_1\lambda_i^1\lambda_i^2 + \mu_\star^2\eta_i^1\eta_i^2\right)\left[\mu_\star^2\left[\eta^1(G_W^{-1})\eta^2\right] + \mu_1^2\psi_1\left[\lambda^1 H_W\lambda^2\right] + 2\mu_1^2\mu_\star^2\psi_1[\lambda^1 S_W\eta^2]\right]$$
$$\exp\left(-\frac{n}{2}\log\det(G_X) - \frac{n}{2}\log\det(G_W) - \frac{1}{2}\psi_1\psi_2\lambda\sum(\eta_h^\alpha)^2 + i\hat{\lambda}_i^\alpha(\sqrt{P}\lambda_i^\alpha - \eta_h^\alpha\boldsymbol{\Theta}_{hi})\right),$$

with

$$(H_W)_{ij} = (G_X^{-1})_{ij} + \frac{\mu_1^2\mu_\star^2\psi_1}{D^2}\left[\eta^\alpha(G_W^{-1})\eta^\beta\right]\left[G_X^{-1}\lambda^\alpha\right]_i\left[G_X^{-1}\lambda^\beta\right]_j, \tag{32}$$

$$(S_W)_{ih} = \frac{1}{D}\left[G_X^{-1}\lambda^\alpha\right]_i\left[G_W^{-1}\eta^\alpha\right]_h. \tag{33}$$

3.3.3 Averaging over the random feature vectors

The expectation over the random feature vectors $\boldsymbol{\Theta}_{hi}$ is a Gausssian integral of the form (5) with:

$$[G_\Theta]_{hh',ii'} = \delta_{hh',ii'}, \tag{34}$$

$$[J_\Theta]_{hi} = -i\hat{\lambda}_i^\alpha\eta_h^\alpha. \tag{35}$$

Preforming this integration results in:

$$\Psi_3^v = \frac{\psi_2}{D}\int\left(\prod_{\alpha=1}^{2n} d\eta^\alpha d\lambda^\alpha d\hat{\lambda}^\alpha\right)\left(\mu_1^2\psi_1\lambda_i^1\lambda_i^2 + \mu_\star^2\eta_i^1\eta_i^2\right)\left[\mu_\star^2\left[\eta^1(G_W^{-1})\eta^2\right] + \mu_1^2\psi_1\left[\lambda^1 H_W\lambda^2\right] + 2\mu_1^2\mu_\star^2\psi_1[\lambda^1 S_W\eta^2]\right]$$
$$\exp\left(-\frac{n}{2}\log\det(G_X) - \frac{n}{2}\log\det(G_W) - \frac{1}{2}\psi_1\psi_2\lambda\sum(\eta_h^\alpha)^2 - \frac{1}{2}\eta_h^\alpha\eta_h^\beta\hat{\lambda}_i^\alpha\hat{\lambda}_i^\beta + i\sqrt{P}\hat{\lambda}_i^\alpha\lambda_i^\alpha\right).$$

3.3.4 Expression of the action and the prefactor

To complete the computation we integrate with respect to $\hat{\lambda}_i^\alpha$, using again formulae (5):

$$\left[G_{\hat{\lambda}}\right]_{ii'}^{\alpha\beta} = \delta^{ii'}\eta_h^\alpha\eta_h^\beta, \tag{36}$$

$$\left[J_{\hat{\lambda}}\right]_i^\alpha = i\sqrt{P}\lambda_i^\alpha. \tag{37}$$

This yields the final expression of the term:

$$\Psi_3^v = \frac{\psi_2}{D}\int\left(\prod_{\alpha=1}^{2n} d\eta^\alpha\right)\left(\prod_{\alpha=1}^{2n} d\lambda^\alpha\right)\left(\mu_1^2\psi_1\lambda_i^1\lambda_i^2 + \mu_\star^2\eta_i^1\eta_i^2\right)\left[\mu_\star^2\left[\eta^1(G_W^{-1})\eta^2\right] + \mu_1^2\psi_1\left[\lambda^1 H_W\lambda^2\right] + 2\mu_1^2\mu_\star^2\psi_1[\lambda^1 S_W\eta^2]\right]$$
$$\exp\left(-\frac{n}{2}\log\det(G_X) - \frac{n}{2}\log\det(G_W) - \frac{D}{2}\log\det(G_{\hat{\lambda}}) - \frac{1}{2}\psi_1\psi_2\lambda\sum(\eta_h^\alpha)^2 - \frac{P}{2}\lambda_i^\alpha(G_{\hat{\lambda}}^{-1})_{ii'}^{\alpha\beta}\lambda_{i'}^\beta\right).$$

The above may be written as

$$\Psi_3^v = \int \left(\prod d\eta\right)\left(\prod d\lambda\right) P_{\Psi_3^v}[\eta,\lambda] \exp\left(-\frac{D}{2} S^v[\eta,\lambda]\right), \tag{38}$$

with the *prefactor* $P_{\Psi_3^v}$ and the *action* S^v defined as:

$$\begin{aligned} P_{\Psi_3^v}[\eta,\lambda] :=& \frac{\psi_2}{D}\left(\mu_1^2\psi_1\lambda_i^1\lambda_i^2 + \mu_\star^2\eta_i^1\eta_i^2\right)\left[\mu_\star^2\left[\eta^1(G_W^{-1})\eta^2\right] + \mu_1^2\psi_1\left[\lambda^1 H_W \lambda^2\right] + 2\mu_1^2\mu_\star^2\psi_1[\lambda^1 S_W \eta^2]\right], \\ S^v[\eta,\lambda] :=& \psi_2 \log\det(G_X) + \psi_2\log\det(G_W) + \log\det(G_{\hat{\lambda}}) + \frac{1}{D}\psi_1\psi_2\lambda\sum(\eta_h^\alpha)^2 + \frac{P}{D}\left(\lambda_i^\alpha (G_{\hat{\lambda}}^{-1})_{ii'}^{\alpha\beta}\lambda_{i'}^\beta\right). \end{aligned}$$

3.3.5 Expression of the action and the prefactor in terms of order parameters

Here we see that we have a factor $D \to \infty$ in the exponential part, which can be estimated using the saddle point method. Before doing so, we introduce the following order parameters using the Fourier representation of the delta-function:

$$1 = \int dQ_{\alpha\beta} d\hat{Q}_{\alpha\beta} e^{\hat{Q}_{\alpha\beta}(PQ_{\alpha\beta} - \eta_h^\alpha \eta_h^\beta)}, \tag{39}$$

$$1 = \int dR_{\alpha\beta} d\hat{R}_{\alpha\beta} e^{\hat{R}_{\alpha\beta}(DR_{\alpha\beta} - \lambda_i^\alpha \lambda_i^\beta)}. \tag{40}$$

This allows to rewrite the prefactor only in terms of Q, R: for example,

$$\mu_1^2\psi_1\lambda_i^1\lambda_i^2 + \mu_\star^2\eta_i^1\eta_i^2 = \psi_1 D(\mu_1^2 R^{12} + \mu_\star^2 Q^{12}).$$

To do this, there are two key quantities we need to calculate: $\lambda G_X^{-1}\lambda$ and $\eta G_W^{-1}\eta$. To calculate both, we note that G_X ang G_W are both of the form $\mathbf{I} + \mathbf{X}$, therefore there inverse may be calculated using their series representation. The result is:

$$[M_X]^{\alpha\beta} := \frac{1}{D}\lambda^\alpha G_X^{-1}\lambda^\beta = \left[R(I + \mu_1^2\psi_1 R)^{-1}\right]^{\alpha\beta}, \tag{41}$$

$$[M_W]^{\alpha\beta} := \frac{1}{P}\eta^\alpha (G_W^{-1})\eta^\beta = \left[Q(I + \mu_\star^2\psi_1 AQ)^{-1}\right]^{\alpha\beta}. \tag{42}$$

Using the above, we deduce:

$$\lambda^1 H_W \lambda^2 = DM_X^{12} + P\mu_1^2\mu_\star^2\psi_1\left[M_X M_W M_X\right]^{12}, \tag{43}$$

$$\lambda^1 S_W \eta^2 = P\left[M_X M_W\right]^{12}. \tag{44}$$

The integrals over η, λ become simple Gaussian integrals with covariance matrices given by $\hat{Q}, \hat{R}$, yielding:

$$1 = \int dQ_{\alpha\beta} d\hat{Q}_{\alpha\beta} e^{-\frac{\psi_1 D}{2}\left(\log\det\hat{Q} - 2\mathrm{Tr}Q\hat{Q}\right)}, \tag{45}$$

$$1 = \int dR_{\alpha\beta} d\hat{R}_{\alpha\beta} e^{-\frac{D}{2}\left(\log\det\hat{R} - 2\mathrm{Tr}R\hat{R}\right)}. \tag{46}$$

The next step is to take the saddle point with respect to the auxiliary variables $\hat{Q}$ and $\hat{R}$ in order to eliminate them:

$$\frac{\partial S^v}{\partial \hat{Q}_{\alpha\beta}} = \psi_1\left(\hat{Q}^{-1} - 2Q\right) = 0 \Rightarrow \hat{Q} = \frac{1}{2}Q^{-1}, \tag{47}$$

$$\frac{\partial S^v}{\partial \hat{R}_{\alpha\beta}} = \left(\hat{R}^{-1} - 2R\right) = 0 \Rightarrow \hat{R} = \frac{1}{2}R^{-1}. \tag{48}$$

One finally obtains that:

$$\Psi_3^v = \int \left(\prod dQ\right)\left(\prod dR\right) P_{\Psi_3^v}[Q,R] \exp\left(-\frac{D}{2} S^v[Q,R]\right), \tag{49}$$

With:

$$\begin{aligned} P_{\Psi_3^v}[Q,R] =& D\psi_1^2\psi_2(\mu_1^2 R^{12} + \mu_\star^2 Q^{12})\left[\mu_\star^2\left[M_W^v\right]^{12} + \mu_1^2\left(M_X^{12} + \mu_1^2\mu_\star^2\psi_1^2\left[M_X M_W^v M_X\right]^{12}\right) + 2\mu_1^2\mu_\star^2\psi_1\left[M_X M_W^v\right]^{12}\right], \\ S^v[Q,R] =& \psi_2\log\det(G_X) + \psi_2\log\det(G_W) + \psi_1^2\psi_2\lambda\mathrm{Tr}Q + \mathrm{Tr}\left(RQ^{-1}\right) + (1-\psi_1)\log\det Q - \log\det R. \end{aligned} \tag{50}$$

3.3.6 Saddle point equations

The aim is now to use the saddle point method in order to evaluate the integrals over the order parameters. Thus, one looks for R and Q solutions to the equations:

$$\frac{\partial S^v}{\partial Q_{\alpha\beta}} = 0, \quad \frac{\partial S^v}{\partial R_{\alpha\beta}} = 0 \quad \forall \alpha, \beta = 1, \cdots, 2n.$$

To solve the above, it is common to make a *replica symmetric ansatz*. In this case, we assume that the solutions to the saddle points equations take the form:

$$Q = \begin{bmatrix} q & \tilde{q} & \cdots & \tilde{q} \\ \tilde{q} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \tilde{q} \\ \tilde{q} & \cdots & \tilde{q} & q \end{bmatrix}, \quad R = \begin{bmatrix} r & \tilde{r} & \cdots & \tilde{r} \\ \tilde{r} & \ddots & \ddots & \vdots \\ \vdots & \ddots & \ddots & \tilde{r} \\ \tilde{r} & \cdots & \tilde{r} & r \end{bmatrix} \tag{51}$$

The action takes the following form:

$$\begin{aligned}
S^v(q, r, \tilde{q}, \tilde{r}) &= 2n \left(S_0(q, r) + S_1^v(q, r, \tilde{q}, \tilde{r}) \right) \\
S_0(q, r) &= \lambda \psi_1^2 \psi_2 q + \psi_2 \log \left(\frac{\mu_\star^2 \psi_1 q}{\mu_1^2 \psi_1 r + 1} + 1 \right) + \frac{r}{q} + (1 - \psi_1) \log(q) + \psi_2 \log \left(\mu_1^2 \psi_1 r + 1 \right) - \log(r) \\
S_1^v(q, r, \tilde{q}, \tilde{r}) &= f^v(q, r) \tilde{q} + g^v(q, r) \tilde{r} \\
f^v(q, r) &= \lambda \psi_1^2 \psi_2 + \frac{\mu_\star^2 \psi_1 \psi_2}{\mu_\star^2 \psi_1 q + \mu_1^2 \psi_1 r + 1} + \frac{1 - \psi_1}{q} - \frac{r}{q^2} \\
g^v(q, r) &= -\frac{\mu_\star^2 \mu_1^2 \psi_1^2 \psi_2 q}{\left(\mu_1^2 \psi_1 r + 1 \right) \left(\mu_\star^2 \psi_1 q + \mu_1^2 \psi_1 r + 1 \right)} + \frac{\mu_1^2 \psi_1 \psi_2}{\mu_1^2 \psi_1 r + 1} + \frac{1}{q} - \frac{1}{r}.
\end{aligned} \tag{52}$$

3.3.7 Fluctuations around the saddle point

We introduce the following notations:

$$\begin{aligned}
\left[\nabla_{\boldsymbol{T}} F(\boldsymbol{T}_\star) \right]_{\alpha\beta} &= \frac{\partial F}{\partial \boldsymbol{T}_{\alpha\beta}} |_{\boldsymbol{T}_\star}, \\
\left[H_{\boldsymbol{T}} F(\boldsymbol{T}_\star) \right]_{\alpha\beta, \gamma\delta} &= \left[\begin{matrix} \frac{\partial^2 F}{\partial Q_{\alpha\beta} \partial Q_{\gamma\delta}} & \frac{\partial^2 F}{\partial Q_{\alpha\beta} \partial R_{\gamma\delta}} \\ \frac{\partial^2 F}{\partial R_{\alpha\beta} \partial Q_{\gamma\delta}} & \frac{\partial^2 F}{\partial R_{\alpha\beta} \partial R_{\gamma\delta}} \end{matrix} \right] \Bigg|_{\boldsymbol{T}_\star}, \\
H\left[F\right] &= \left[\begin{matrix} \frac{\partial F}{\partial q \partial q} & \frac{\partial F}{\partial q \partial r} & \frac{\partial F}{\partial q \partial \tilde{q}} & \frac{\partial F}{\partial q \partial \tilde{r}} \\ \frac{\partial F}{\partial q \partial r} & \frac{\partial F}{\partial r \partial r} & \frac{\partial F}{\partial r \partial \tilde{q}} & \frac{\partial F}{\partial r \partial \tilde{r}} \\ \frac{\partial F}{\partial q \partial \tilde{q}} & \frac{\partial F}{\partial r \partial \tilde{q}} & \frac{\partial F}{\partial \tilde{q} \partial \tilde{q}} & \frac{\partial F}{\partial \tilde{q} \partial \tilde{r}} \\ \frac{\partial F}{\partial q \partial \tilde{r}} & \frac{\partial F}{\partial r \partial \tilde{r}} & \frac{\partial F}{\partial \tilde{q} \partial \tilde{r}} & \frac{\partial F}{\partial \tilde{r} \partial \tilde{r}} \end{matrix} \right] \Bigg|_{\substack{q=q^* \\ r=r^* \\ \tilde{r}=0 \\ \tilde{q}=0}}.
\end{aligned}$$

Proposition Let $q^\star$ and $r^\star$ be the solutions of the fixed point equation for the function $S_0 : (q, r) \mapsto \mathbb{R}$ defined in (52):

$$\begin{cases} \frac{\partial S_0(q,r)}{\partial q} = 0 \\ \frac{\partial S_0(q,r)}{\partial r} = 0. \end{cases}$$

Then we have that

$$\Psi_3^v = \frac{1}{D} \operatorname{Tr} \left[H\left[S^v\right]^{-1} H\left[P_{\Psi_3^v}\right] \right]. \tag{53}$$

Sketch of proof Solving the saddle point equations:

$$\begin{cases} \frac{\partial S^v(q,r,\tilde{q},\tilde{r})}{\partial q} = 0 \\ \frac{\partial S^v(q,r,\tilde{q},\tilde{r})}{\partial q} = 0 \\ \frac{\partial S^v(q,r,\tilde{q},\tilde{r})}{\partial \tilde{q}} = 0 \\ \frac{\partial S^v(q,r,\tilde{q},\tilde{r})}{\partial \tilde{r}} = 0 \end{cases},$$

one finds $\tilde{q} = \tilde{r} = 0$, which is problematic because the prefactor vanishes: $P_{\Psi_3} \propto \mu_1^2\tilde{q} + \mu_\star^2\tilde{r}$.
Therefore we must go beyond the saddle point contribution to obtain a non zero result, i.e. we have to examine the quadratic fluctuations around the saddle point. To do so we preform a second-order expansion of the action (50) as a function of Q and R:

$$\begin{aligned} P_{\Psi_3^v}(\boldsymbol{T}) &\approx P_{\Psi_3^v}(\boldsymbol{T}_\star) + (\boldsymbol{T}-\boldsymbol{T}_\star)^\top \nabla P_{\Psi_3^v}(\boldsymbol{T}_\star) + \frac{1}{2}(\boldsymbol{T}-\boldsymbol{T}_\star)^\top H_{\boldsymbol{T}}\left[P_{\Psi_3^v}(\boldsymbol{T}_\star)\right](\boldsymbol{T}-\boldsymbol{T}_\star), \\ S^v(\boldsymbol{T}) &\approx S_0(\boldsymbol{T}_\star) + \frac{1}{2}(\boldsymbol{T}-\boldsymbol{T}_\star)^\top H_{\boldsymbol{T}}\left[S^v(\boldsymbol{T}_\star)\right](\boldsymbol{T}-\boldsymbol{T}_\star). \end{aligned}$$

Computing the second derivative of (50), it is easy to show that:

$$\begin{aligned} \left[H_{\boldsymbol{T}}\left[S^v(\boldsymbol{T})\right]\right]_{\alpha\beta,\gamma\delta} &= \left[H_{\boldsymbol{T}}\left[S^v(\boldsymbol{T})\right]\right]_{\alpha\beta}\left(\delta_{\alpha\gamma}\delta_{\beta\delta} + \delta_{\alpha\delta}\delta_{\beta\gamma}\right), \\ \left[H_{\boldsymbol{T}}\left[P_{\Psi_3^v}(\boldsymbol{T})\right]\right]_{\alpha\beta,\gamma\delta} &= \left[H_{\boldsymbol{T}}\left[P_{\Psi_3^v}(\boldsymbol{T})\right]\right]_{\alpha\beta}\left(\delta_{\alpha\gamma}\delta_{\beta\delta} + \delta_{\alpha\delta}\delta_{\beta\gamma}\right), \end{aligned}$$

where

$$\begin{aligned} H_{\boldsymbol{T}}[F]_{\alpha\beta} &= \begin{bmatrix} \frac{\partial^2 F}{\partial Q_{\alpha\beta}\partial Q_{\alpha\beta}} & \frac{\partial^2 F}{\partial Q_{\alpha\beta}\partial R_{\alpha\beta}} \\ \frac{\partial^2 F}{\partial R_{\alpha\beta}\partial Q_{\alpha\beta}} & \frac{\partial^2 F}{\partial R_{\alpha\beta}\partial R_{\alpha\beta}} \end{bmatrix} \\ &= \frac{1}{2n}\delta_{\alpha\beta}\begin{bmatrix} \frac{\partial^2 F}{\partial q\partial q} & \frac{\partial^2 F}{\partial q\partial r} \\ \frac{\partial^2 F}{\partial r\partial q} & \frac{\partial^2 F}{\partial r\partial r} \end{bmatrix} + \frac{2}{2n(2n-1)}(1-\delta_{\alpha\beta})\begin{bmatrix} \frac{\partial^2 F}{\partial \tilde{q}\partial \tilde{q}} & \frac{\partial^2 F}{\partial \tilde{q}\partial \tilde{r}} \\ \frac{\partial^2 F}{\partial \tilde{r}\partial \tilde{q}} & \frac{\partial^2 F}{\partial \tilde{r}\partial \tilde{r}} \end{bmatrix}. \end{aligned}$$

Hence,

$$\begin{aligned} \Psi_3^v &= \lim_{n\to 0}\int d\boldsymbol{T} P_{\Psi_3^v}(\boldsymbol{T}) \exp^{-\frac{D}{2}S^v(\boldsymbol{T})}, \\ &= \lim_{n\to 0} \underbrace{P_{\Psi_3^v}(\boldsymbol{T}_\star)}_{0} + \nabla P_{\Psi_3^v}(\boldsymbol{T}_\star)_{\alpha\beta} \underbrace{\int d\boldsymbol{T}(\boldsymbol{T}-\boldsymbol{T}_\star)_{\alpha\beta}\exp\left(-\frac{D}{2}S^v(\boldsymbol{T})\right)}_{0} \\ &\quad + \frac{1}{2}H_{\boldsymbol{T}}\left[P_{\Psi_3^v}(\boldsymbol{T}_\star)\right]_{\alpha\beta}\int d\boldsymbol{T}(\boldsymbol{T}-\boldsymbol{T}_\star)^2_{\alpha\beta}\exp\left(-\frac{D}{2}S^v(\boldsymbol{T})\right) \\ &= \lim_{n\to 0}\frac{1}{2}H_{\boldsymbol{T}}\left[P_{\Psi_3^v}(\boldsymbol{T}_\star)\right]_{\alpha\beta} e^{-\frac{D}{2}S^v(\boldsymbol{T}_\star)}\int d\boldsymbol{T}(\boldsymbol{T}-\boldsymbol{T}_\star)^2_{\alpha\beta} e^{-\frac{D}{2}\sum_{\alpha\beta}(\boldsymbol{T}-\boldsymbol{T}_\star)^2_{\alpha\beta}H_{\boldsymbol{T}}[S^v(\boldsymbol{T}_\star)]_{\alpha\beta}} \\ &= \lim_{n\to 0}\frac{1}{2}\frac{(2\pi)^n}{\det H_{\boldsymbol{T}}\left[S^v(\boldsymbol{T}_\star)\right]} e^{-\frac{D}{2}S^v(\boldsymbol{T}_\star)}\frac{2}{D}H_{\boldsymbol{T}}\left[P_{\Psi_3^v}(\boldsymbol{T}_\star)\right]_{\alpha\beta} H_{\boldsymbol{T}}\left[S^v(\boldsymbol{T}_\star)\right]^{-1}_{\alpha\beta} \\ &= \frac{1}{D}\operatorname{Tr}\left[H\left[S^v\right]^{-1} H\left[P_{\Psi_3^v}\right]\right] \end{aligned}$$

In the last step, we used the fact that:

$$\begin{aligned} \lim_{n\to 0} e^{-\frac{D}{2}S^v(\boldsymbol{T}_\star)} &= \lim_{n\to 0} e^{-nDS_0(q_\star,r_\star)} = 1, \\ \lim_{n\to 0}\det H_{\boldsymbol{T}}\left[S^v(\boldsymbol{T}_\star)\right] &= 1. \end{aligned}$$

The last equality follows from the fact that for a matrix of size $n\times n$ of the form $M_{\alpha\beta} = a\delta_{\alpha\beta} + b(1-\delta_{\alpha\beta})$, we have

$$\det M = (a-b)^n\left(1 + \frac{nb}{a-b}\right) \xrightarrow[n\to 0]{} 1.$$

3.3.8 Expression of the vanilla terms

Using the above procedure, one can compute the terms $\Psi_1, \Psi_2^v, \Psi_3^v$ of (24): for each of these terms, the action is the same as in (50), and the prefactors can be obtained as:

$$P_{\Psi_1}[Q,R] = \mu_1^2\psi_1\psi_2\left(M_X^{11} + \mu_1^2\mu_\star^2\psi_1^2\left(M_X M_W M_X\right)^{11} + \mu_\star^2\mu_1\psi_1\left(M_X M_W\right)^{11}\right),$$
$$P_{\Psi_2^v}[Q,R] = D\psi_1^2\psi_2\left(\mu_1^2R^{12} + \mu_\star^2Q^{12}\right)\left(\mu_1^2\psi_2 P_{XX} - 2\mu_1^2\mu_\star^2\psi_1\psi_2 P_{XW} + \mu_\star^2 P_{WW}\right),$$
$$P_{\Psi_3^v}[Q,R] = D\psi_1^2\psi_2(\mu_1^2R^{12} + \mu_\star^2Q^{12})\left[\mu_\star^2\left[M_W^v\right]^{12} + \mu_1^2\left(M_X^{12} + \mu_1^2\mu_\star^2\psi_1^2\left[M_X M_W^v M_X\right]^{12}\right) + 2\mu_1^2\mu_\star^2\psi_1\left[M_X M_W^v\right]^{12}\right],$$
$$P_{XX} = N_X^{12} + \frac{1}{\psi_2}M_X^{12} + 2\left(\mu_1\mu_\star\psi_1\right)^2\left[N_X M_W M_X\right]^{12} + \frac{\left(\mu_1\mu_\star\psi_1\right)^2}{\psi_2}[M_X M_W M_X]^{12} + \left(\mu_1\mu_\star\psi_1\right)^4\left[M_X M_W N_X M_W M_X\right]^{12},$$
$$P_{XW} = \left[N_X M_W\right]^{12} + \frac{1}{\psi_2}\left[M_X M_W\right]^{12} + \left(\mu_1\mu_\star\psi_1\right)^2\left[M_X M_W N_X M_W\right]^{12},$$
$$P_{WW} = M_W^{12} + \psi_2\left(\mu_1\mu_\star\psi_1\right)^2\left[M_W N_X M_W\right]^{12}.$$

Where a new term appears:

$$[N_X]^{\alpha\beta} = \left[R(I + \mu_1^2\psi_1 R)^{-2}\right]^{\alpha\beta}. \tag{54}$$

3.4 Computation of the *ensembling* terms

3.4.1 Expression of the action and the prefactor

In the *ensembling* terms, the two inverse matrices are different, hence one has to introduce two distinct replica variables. We distinguish them by the use of an extra index $a \in \{1,2\}$, denoted in brackets in order not to be confused with the replica indices α.

$$\left[M^{(1)}\right]^{-1}_{ij}\left[M^{(2)}\right]^{-1}_{kl} = \lim_{n\to 0}\int\left(\prod_{\alpha=1}^{n}\prod_{a=1}^{2}d\eta^{\alpha(a)}\right)\eta_i^{1(1)}\eta_j^{1(1)}\eta_k^{1(2)}\eta_l^{1(2)}\exp\left(-\frac{1}{2}\sum_{(a)}\eta^{\alpha(a)}M_{ij}^{(a)}\eta^{\alpha(a)}\right). \tag{55}$$

Calculations of the Gaussian integrals follow through in a very similar way as for the *vanilla* terms. The matrices appearing in the process are:

$$(G_X^e)_{ii'} = \delta_{ii'} + \frac{\mu_1^2\psi_1}{D}\sum_{(a)\alpha}\lambda_i^{\alpha(a)}\lambda_{i'}^{\alpha(a)}, \tag{56}$$
$$(J_X^e)_i = \frac{\mu_1\mu_\star\sqrt{\psi_1}}{D^2}\sum_{\alpha a}\lambda_i^{\alpha(a)}\eta_h^{\alpha(a)}W_h^{(a)}, \tag{57}$$
$$(G_W^e)_{hh'}^{(ab)} = \delta_{hh'}^{ab} + \frac{\mu_\star^2}{D}\sum_{\alpha\beta}\eta_h^{\alpha(a)}A_e^{\alpha\beta,ab}\eta_{h'}^{\beta(b)}, \tag{58}$$
$$(J_W^e)_h = 0 \tag{59}$$
$$(G_\Theta^e)_{hh',ii'}^{(ab)} = \delta_{hh',ii'}^{ab} \tag{60}$$
$$(J_\Theta^e)_{hi}^{(a)} = -i\sum_{\alpha}\hat{\lambda}_i^{\alpha(a)}\eta_h^{\alpha(a)}, \tag{61}$$
$$(G_{\hat{\lambda}}^e)_{ii'}^{\alpha\beta,(ab)} = 2\delta^{ab}\delta^{ii'}\eta_h^{\alpha(a)}\eta_h^{\beta(b)}, \tag{62}$$
$$(J_{\hat{\lambda}}^e)_i^{\alpha(a)} = i\sqrt{P}\lambda_i^{\alpha(a)}. \tag{63}$$
$$\tag{64}$$

with

$$A_e^{\alpha\beta,(ab)} = \delta_{ab}\delta_{\alpha\beta} - \mu_1^2\psi_1 \frac{1}{D}\sum_{i,j} \lambda_i^{\alpha(a)} [G_X^{e-1}]_{ij} \lambda_j^{\beta(b)}, \tag{65}$$

$$[H_W^e]_{ii'} = \left[G_X^{e-1}\right]_{ii'} + \sum_{\alpha\beta,ab} \left[G_X^{e-1}\lambda^{\alpha(a)}\right]_i \left[G_X^{e-1}\lambda^{\beta(b)}\right]_{i'} \left[\eta^{\alpha(a)} \left[G_W^{e-1}\right]^{(ab)} \eta^{\beta(b)}\right], \tag{66}$$

$$[S_W^e]_{ih} = \frac{1}{D}\sum_{\alpha,a} \left[G_X^{e-1}\lambda^{\alpha(a)}\right]_i \left[G_W^{e-1}\eta^{\alpha(a)}\right]_h. \tag{67}$$

Starting with the computation of Ψ_3^e in order to illustrate the method used, the prefactor $P_{\Psi_3^e}$ and the action are S^e are given by:

$$P_{\Psi_3^e}[\eta,\lambda] := \frac{\mu_1^2\psi_1\psi_2}{D}\lambda_i^{1(1)}\lambda_i^{1(2)} \left[\mu_\star^2 \left[\eta^{1(1)}(G_W^{e-1})^{(12)}\eta^{1(2)}\right] + \mu_1^2\psi_1 \left[\lambda^{1(1)}H_W^e\lambda^{1(2)}\right] + 2\mu_1^2\mu_\star^2\psi_1[\lambda^{1(1)}S_W^e\eta^{1(2)}]\right], \tag{68}$$

$$S^e[\eta,\lambda] := \psi_2\log\det(G_X^e) + \psi_2\log\det(G_W^e) + \log\det(G_{\hat\lambda}^e) + \frac{1}{D}\psi_1\psi_2\lambda\sum(\eta_h^{\alpha(a)})^2 + \frac{1}{2D}\left(\lambda_i^{\alpha(a)}(G_{\hat\lambda}^{e-1})_{ii'}^{\alpha\beta,ab}\lambda_{i'}^{\beta(b)}\right). \tag{69}$$

3.4.2 Expression of the action and the prefactor in terms of order parameters

This time, because of the two different systems, the order parameters carry an additional index a, which turns them into 2×2 block matrices:

$$1 = \int dQ_{\alpha\beta}^{(ab)} d\hat{Q}_{\alpha\beta}^{(ab)} e^{\hat{Q}_{\alpha\beta}^{(ab)}(PQ_{\alpha\beta}^{(ab)} - \eta_h^{\alpha(a)}\eta_h^{\beta(b)})}, \tag{70}$$

$$1 = \int dR_{\alpha\beta}^{(ab)} d\hat{R}_{\alpha\beta}^{(ab)} e^{\hat{R}_{\alpha\beta}^{(ab)}(dR_{\alpha\beta}^{(ab)} - \lambda_i^{\alpha(a)}\lambda_i^{\beta(b)})}. \tag{71}$$

The systems being decoupled, we make the following ansatz for the order parameters:

$$Q = \left[\begin{array}{ccc|ccc} q & \cdots & 0 & \tilde{q} & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & q & 0 & \cdots & \tilde{q} \\ \hline \tilde{q} & \cdots & 0 & q & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \tilde{q} & 0 & \cdots & q \end{array}\right], \quad R = \left[\begin{array}{ccc|ccc} r & \cdots & 0 & \tilde{r} & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & r & 0 & \cdots & \tilde{r} \\ \hline \tilde{r} & \cdots & 0 & r & \cdots & 0 \\ \vdots & \ddots & \vdots & \vdots & \ddots & \vdots \\ 0 & \cdots & \tilde{r} & 0 & \cdots & r \end{array}\right]. \tag{72}$$

In virtue of the simple structure of the above matrices, the replica indices α trivialize and we may replace the matrices Q and R by the 2×2 matrices:

$$Q = \begin{bmatrix} q & \tilde{q} \\ \tilde{q} & q \end{bmatrix}, \quad R = \begin{bmatrix} r & \tilde{r} \\ \tilde{r} & r \end{bmatrix}.$$

Define:

$$[M_X^e]_{(ab)} \equiv \frac{1}{D}\lambda_i^{(a)}\left[G_X^{e-1}\right]_{ij}\lambda_j^{(b)} = \left[R(I+\mu_1^2\psi_1 R)^{-1}\right]_{(ab)}, \tag{73}$$

$$[M_W^e]_{(ab)} \equiv \frac{1}{P}\eta^{(a)}\left[G_W^{e-1}\right]^{(ab)}\eta^{(b)} = \left[Q(I+\psi_1 A^e Q)^{-1}\right]_{(ab)}, \tag{74}$$

where products are now over 2×2 matrices. Then, one has:

$$P_{\Psi_3^e}[Q,R] := D\mu_1^2\psi_1^2\psi_2 R^{(12)} \left[\mu_\star^2 M_W^{e(12)} + \mu_1^2\left(M_X^{e(12)} + \mu_1^2\mu_\star^2\psi_1^2\left[M_X^e M_W^e M_X^e\right]^{(12)}\right) + 2\mu_1^2\mu_\star^2\psi_1\left[M_X^e M_W^e\right]^{(12)}\right],$$

$$S^e[Q,R] := \psi_2\log\det(G_X^e) + \psi_2\log\det(G_W^e) + \sum_a \left[(\psi_1^2\psi_2\lambda)\mathrm{Tr}Q^{(a)(a)} + \mathrm{Tr}\left(R^{(a)(a)}(Q^{-1})^{(a)(a)}\right) + \log\det Q^{(a)(a)}\right]$$

$$- \psi_1\log\det Q - \log\det R.$$

Where we have:

$$M_X^e = \frac{1}{(1+\mu_1^2\psi_1 r)^2 - (\mu_1^2\psi_1\tilde{r})^2}\begin{bmatrix} r+\mu_1^2\psi_1(r^2-\tilde{r}^2) & \tilde{r} \\ \tilde{r} & r+\mu_1^2\psi_1(r^2-\tilde{r}^2)\end{bmatrix}, \tag{75}$$

$$A^e_{(ab)} = \delta_{(ab)} - \mu_1^2\psi_1\left[M_X^e\right]_{(ab)}, \tag{76}$$

$$\left[M_W^e\right]_{(ab)} = q\delta_{(ab)} - \mu_\star^2\psi_1\left[A^e(I+\mu_\star^2\psi_1 q A^e)^{-1}\right]_{(ab)}, \tag{77}$$

$$\det(G_X^e) = \det(\delta_{(ab)} + \psi_1\mu_1^2 R_{(ab)}), \tag{78}$$

$$\det(G_W^e) = \det(\delta_{ab} + \psi_1 q A^e_{(ab)}). \tag{79}$$

Finally, we are left with:

$$\begin{aligned} S^e(q,r,\tilde{q},\tilde{r}) =& n\left(S_0(q,r) + S_1^e(q,r,\tilde{q},\tilde{r})\right), \\ S_1^e(q,r,\tilde{q},\tilde{r}) =& \tilde{r}^2 f^e(q,r) + \tilde{q}^2 g^e(q,r), \\ f^e(q,r) =& \frac{2r\mu_1^2\psi_1\left(1+q\mu_\star^2\psi_1\right) + \left(1+q\mu_\star^2\psi_1\right)^2 - r^2\mu_1^4\psi_1^2(-1+\psi_2)}{2r^2\left(1+r\mu_1^2\psi_1+q\mu_\star^2\psi_1\right)^2}, \\ g^e(q,r) =& \frac{\psi_1}{2q^2}. \end{aligned}$$

Where S_0 was defined in (52).

3.4.3 Expression of the ensembling terms

Evaluating the fluctuations around the saddle point follows through in the same way as for the vanilla terms, with the following expressions of the prefactors:

$$P_{\Psi_2^e}[Q,R] = D\psi_1^2\psi_2\mu_1^2\tilde{r}\left[\mu_1^2 P^e_{XX} - 2\mu_1^2\mu_\star^2\psi_1 P^e_{WX} + \mu_\star^2 P^e_{WW}\right], \tag{80}$$

$$P_{\Psi_3^e}[Q,R] = D\mu_1^2\psi_1^2\psi_2 R^{(12)}\left[\mu_\star^2 M_W^{e(12)} + \mu_1^2\left(M_X^{e(12)} + \mu_1^2\mu_\star^2\psi_1^2\left[M_X^e M_W^e M_X^e\right]^{(12)}\right) + 2\mu_1^2\mu_\star^2\psi_1\left[M_X^e M_W^e\right]^{(12)}\right], \tag{81}$$

$$\begin{aligned} P^e_{XX} =& \psi_2 N_X^{e12} + M_X^{e12} + 2\psi_2(\mu_1\mu_\star\psi_1)^2\left[M_X^e N_X^e M_W^e\right]^{12} + (\mu_1\mu_\star\psi_1)^2\left[M_X^e M_W^e M_X^e\right]^{12} \\ &+ \psi_2(\mu_1\mu_\star\psi_1)^4\left[M_X^e M_W^e N_X^e M_W^e M_X^e\right]^{12}, \end{aligned} \tag{82}$$

$$P^e_{WX} = \psi_2\left[N_X^e M_W^e\right]^{12} + \left[M_X^e M_W^e\right]^{12} + \psi_2(\mu_1\mu_\star\psi_1)^2\left[M_X^e M_W^e N_X^e M_W^e\right]^{12}, \tag{83}$$

$$P^e_{WW} = \left[M_W^e\right]^{12} + \psi_2(\mu_1\mu_\star\psi_1)^2\left[M_W^e N_X^e M_W^e\right]^{12}. \tag{84}$$

3.5 Computation of the *divide and conquer* term

Here, we are interested in computing the term Ψ_2^d. This term differs from the previous ones in that there are now two independent data matrices $\boldsymbol{X}^{(1)}$ and $\boldsymbol{X}^{(2)}$. The calculations for the action and the prefactor are very similar to calculations performed for the *ensembling* terms Ψ_2^e, Ψ_3^e, with the addition that $\boldsymbol{X}$ now also carries an index (a). Firstly let us write Ψ_2^d as a trace over random matrices:

$$\Psi_2^d = \frac{\mu_1^2}{d^2}\,\mathrm{Tr}\left[\boldsymbol{X}^{(1)}\boldsymbol{Z}^{(1)}\boldsymbol{B}^{(1)-1}\boldsymbol{\Theta}^{(1)}\boldsymbol{\Theta}^{(2)}\boldsymbol{B}^{(2)-1}\boldsymbol{Z}^{(2)}\boldsymbol{X}^{(2)}\right]. \tag{85}$$

Calculations follow through in the same way as in the previous sections. Using the replica formula (55), and performing the integrals over the Gaussian variables the following quantities appear:

$$(G_X^d)_{ii'}^{(ab)} = \delta_{ii'} + \frac{\mu_1^2\psi_1}{D}\sum_\alpha \lambda_i^{\alpha(a)}\lambda_{i'}^{\alpha(a)}, \tag{86}$$

$$(J_X^d)_i^{(a)} = \frac{\mu_1\mu_\star\sqrt{\psi_1}}{D^2}\sum_\alpha \lambda_i^{\alpha(a)}\eta_h^{\alpha(a)}W_h^{(a)}, \tag{87}$$

$$(G_W^d)_{hh'}^{(ab)} = \delta_{hh'}^{ab} + \frac{\mu_\star^2}{D}\sum_{\alpha\beta} \eta_h^{\alpha(a)}A^{\alpha\beta,ab}\eta_{h'}^{\beta(b)}, \tag{88}$$

$$(J_W^d)_h = 0, \tag{89}$$

$$(G_\Theta^d)_{hh',ii'}^{(ab)} = \delta_{hh',ii'}^{ab}, \tag{90}$$

$$(J_\Theta^d)_{hi}^{(a)} = -i\sum_\alpha \hat{\lambda}_i^{\alpha(a)}\eta_h^{\alpha(a)}, \tag{91}$$

$$(G_{\hat{\lambda}}^d)_{ii'}^{\alpha\beta,(ab)} = 2\delta^{ab}\delta^{ii'}\eta_h^{\alpha(a)}\eta_h^{\beta(b)}, \tag{92}$$

$$(J_{\hat{\lambda}}^d)_i^{\alpha(a)} = i\sqrt{P}\lambda_i^{\alpha(a)}. \tag{93}$$

$$\tag{94}$$

The saddle point ansatz for Q and R is the same as the one for the ensembling terms (see (72)). The procedure to evaluate Ψ_2^d is also the same as the one for Ψ_2^e except the Hessian is taken with respect to S^d. The final result is given below.

$$\begin{aligned}
P_{\Psi_2^d}[Q,R] =& D\mu_1^2\psi_1\psi_2^2\tilde{r}\left[\psi_1\mu_1^2 P_{XX} + 2\mu_\star^2\mu_1^2\psi_1^2 P_{WX} + \mu_\star^2\psi_1 P_{WW}\right], \\
P_{XX} =& \left(N_X^{11} + 2(\mu_1\mu_\star\psi_1)^2\left[N_X M_W M_X\right]^{11} + (\mu_1\mu_\star\psi_1)^4\left[M_X M_W N_X M_W M_X\right]^{11}\right), \\
P_{WX} =& \left[N_X M_W\right]^{11} + (\mu_1\mu_\star\psi_1)^2\left[M_X M_W N_X M_W\right]^{11}, \\
P_{WW} =& (\mu_1\mu_\star\psi_1)^2\left[M_W N_X M_W\right]^{11}, \\
S^d[q,r,\tilde{q},\tilde{r}] =& n\left(S_0(q,r) + S_1^d(q,r,\tilde{q},\tilde{r})\right), \\
S_1^d(q,r,\tilde{q},\tilde{r}) =& \frac{\tilde{r}^2}{r^2} + \frac{\psi_1\tilde{q}^2}{q^2}.
\end{aligned}$$

With:

$$\begin{aligned}
M_X &= \frac{r}{1+\mu_1^2\psi_1 r}, \\
A &= 1 - \mu_1^2\psi_1 M_X, \\
M_W &= \frac{q}{1+\mu_\star^2\psi_1 q A}, \\
N_X &= \frac{\tilde{r}}{(1+\mu_1^2\psi_1 r)^2}.
\end{aligned}$$

References

[1] Nitesh V Chawla, Thomas E Moore, Lawrence O Hall, Kevin W Bowyer, W Philip Kegelmeyer, and Clayton Springer. Distributed learning with bagging-like performance. *Pattern recognition letters*, 24(1-3):455–471, 2003.

[2] Harris Drucker, Corinna Cortes, Lawrence D Jackel, Yann LeCun, and Vladimir Vapnik. Boosting and other ensemble methods. *Neural Computation*, 6(6):1289–1301, 1994.

[3] Mario Geiger, Arthur Jacot, Stefano Spigler, Franck Gabriel, Levent Sagun, Stéphane d'Ascoli, Giulio Biroli, Clément Hongler, and Matthieu Wyart. Scaling description of generalization with number of parameters in deep learning. *arXiv preprint arXiv:1901.01608*, 2019.

[4] Song Mei and Andrea Montanari. The generalization error of random features regression: Precise asymptotics and double descent curve. *arXiv preprint arXiv:1908.05355*, 2019.

[5] Yuchen Zhang, John Duchi, and Martin Wainwright. Divide and conquer kernel ridge regression: A distributed algorithm with minimax optimal rates. *The Journal of Machine Learning Research*, 16(1):3299–3340, 2015.