

7 APPENDIX

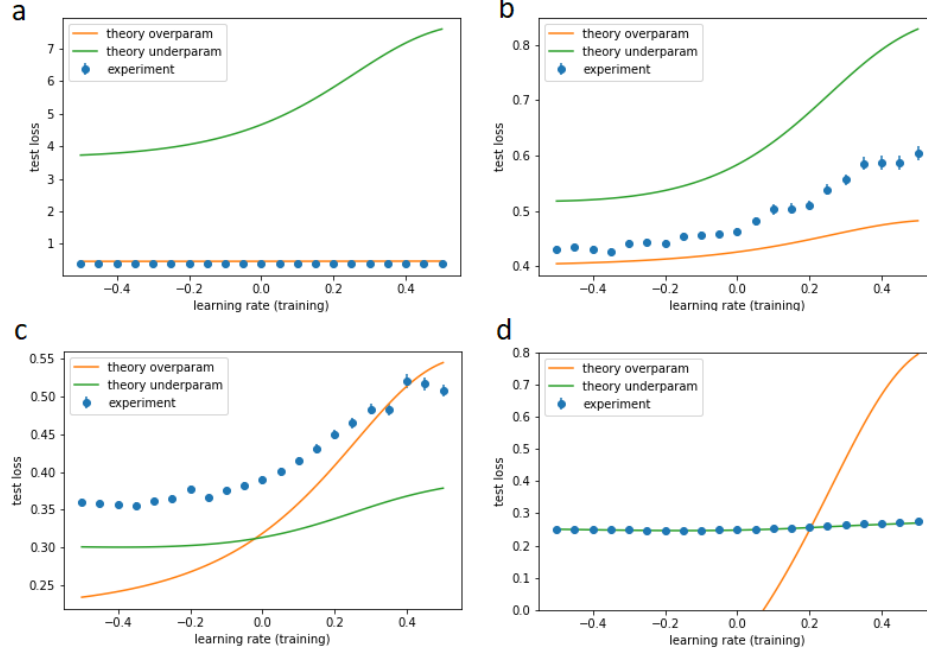


Figure 6: Average test loss of MAML as a function of the learning rate α_t (training) on mixed linear regression, showing the transition from strongly overparameterized (a), to weakly overparameterized (b), weakly underparameterized (c) and strongly underparameterized (d). As expected, predictions of theory are accurate only in panels a and d. The amount of validation data increases from panels a to d, with the following values: $m = 1, n_v = 2$ (a), $m = 5, n_v = 5$ (b), $m = 10, n_v = 10$ (c), $m = 10, n_v = 40$. Other parameters are equal to: $n_t = 40, n_r = 40, p = 50, \sigma = 0.5, \nu = 0.5, \alpha_r = 0.2, \omega_0 = \mathbf{0}, \mathbf{w}_0 = (0.1, 0.1, \dots, 0.1)$ (note that overfitting occurs since $\omega_0 \neq \mathbf{w}_0$). In the experiments, each run is evaluated on 100 tasks of 50 data points each, and each point is an average over 100 runs.

7.1 DEFINITION OF THE LOSS FUNCTION

We consider the problem of mixed linear regression $\mathbf{y} = X\mathbf{w} + \mathbf{z}$ with squared loss, where X is a $n \times p$ matrix of input data, each row is one of n data vectors of dimension p , \mathbf{z} is a $n \times 1$ noise vector, \mathbf{w} is a $p \times 1$ vector of generating parameters and \mathbf{y} is a $n \times 1$ output vector. Data is collected for m tasks, each with a different value of the parameters \mathbf{w} and a different realization of the input X and noise \mathbf{z} . We denote by $\mathbf{w}^{(i)}$ the parameters for task i , for $i = 1, \dots, m$. For a given task i , we denote by $X^{t(i)}, X^{v(i)}$ the input data for, respectively, the training and validation sets, by $\mathbf{z}^{t(i)}, \mathbf{z}^{v(i)}$ the corresponding noise vectors and by $\mathbf{y}^{t(i)}, \mathbf{y}^{v(i)}$ the output vectors. We denote by n_t, n_v the data sample size for training and validations sets, respectively.

For a given task i , the training output is equal to

$$\mathbf{y}^{t(i)} = X^{t(i)}\mathbf{w}^{(i)} + \mathbf{z}^{t(i)} \quad (18)$$

Similarly, the validation output is equal to

$$\mathbf{y}^{v(i)} = X^{v(i)}\mathbf{w}^{(i)} + \mathbf{z}^{v(i)}. \quad (19)$$

We consider MAML as a model for meta-learning (Finn et al 2017). The meta-training loss is equal to

$$\mathcal{L}^{meta} = \frac{1}{2n_v m} \sum_{i=1}^m \left| \mathbf{y}^{v(i)} - X^{v(i)}\boldsymbol{\theta}^{(i)}(\omega) \right|^2 \quad (20)$$

where vertical brackets denote euclidean norm, and the estimated parameters $\theta^{(i)}(\omega)$ are equal to the one-step gradient update on the single-task training loss $\mathcal{L}^{(i)} = \|\mathbf{y}^{t(i)} - X^{t(i)}\theta^{(i)}\|^2/2n_t$, with initial condition given by the meta-parameter ω . The single gradient update is equal to

$$\theta^{(i)}(\omega) = \left(I_p - \frac{\alpha_t}{n_t} X^{t(i)T} X^{t(i)} \right) \omega + \frac{\alpha_t}{n_t} X^{t(i)T} \mathbf{y}^{t(i)} \quad (21)$$

where I_p is the $p \times p$ identity matrix and α_t is the learning rate. We seek to minimize the meta-training loss with respect to the meta-parameter ω , namely

$$\omega^* = \arg \min_{\omega} \mathcal{L}^{meta} \quad (22)$$

We evaluate the solution ω^* by calculating the meta-test loss

$$\mathcal{L}^{test} = \mathbb{E}_{\mathbf{w}'} \mathbb{E}_{\mathbf{z}^s} \mathbb{E}_{X^s} \mathbb{E}_{\mathbf{z}^r} \mathbb{E}_{X^r} \frac{1}{2n_s} \|\mathbf{y}^s - X^s \theta^*\|^2 \quad (23)$$

Note that the test loss is calculated over test data X^s, \mathbf{z}^s , and test parameters \mathbf{w}' , namely

$$\mathbf{y}^s = X^s \mathbf{w}' + \mathbf{z}^s \quad (24)$$

Furthermore, the estimated parameters θ^* are calculated on a separate set of target data X^r, \mathbf{z}^r , namely

$$\theta^* = \left(I_p - \frac{\alpha_r}{n_r} X^{rT} X^r \right) \omega^* + \frac{\alpha_r}{n_r} X^{rT} \mathbf{y}^r \quad (25)$$

$$\mathbf{y}^r = X^r \mathbf{w}' + \mathbf{z}^r \quad (26)$$

Note that the learning rate and sample size can be different at testing, denoted by α_r, n_r, n_s . We are interested in calculating the average test loss, that is averaged over all possible realizations of meta-training data, equal to

$$\bar{\mathcal{L}}^{test} = \mathbb{E}_{\mathbf{w}} \mathbb{E}_{\mathbf{z}^t} \mathbb{E}_{X^t} \mathbb{E}_{\mathbf{z}^v} \mathbb{E}_{X^v} \mathcal{L}^{test} = \mathbb{E}_{\mathbf{w}} \mathbb{E}_{\mathbf{z}^t} \mathbb{E}_{X^t} \mathbb{E}_{\mathbf{z}^v} \mathbb{E}_{X^v} \mathbb{E}_{\mathbf{w}'} \mathbb{E}_{\mathbf{z}^s} \mathbb{E}_{X^s} \mathbb{E}_{\mathbf{z}^r} \mathbb{E}_{X^r} \frac{1}{2n_s} \|\mathbf{y}^s - X^s \theta^*\|^2 \quad (27)$$

7.2 DEFINITION OF PROBABILITY DISTRIBUTIONS

We assume that all random variables are Gaussian. In particular, we assume that the rows of the matrix X are independent, and each row, denoted by \mathbf{x} , is distributed according to a multivariate Gaussian with zero mean and unitary covariance

$$\mathbf{x} \sim \mathcal{N}(0, I_p) \quad (28)$$

where I_p is the $p \times p$ identity matrix. Similarly, the noise is distributed following a multivariate Gaussian with zero mean and variance equal to σ^2 , namely

$$\mathbf{z} \sim \mathcal{N}(0, \sigma^2 I_n) \quad (29)$$

Finally, the generating parameters are also distributed according to a multivariate Gaussian of variance ν^2 , namely

$$\mathbf{w} \sim \mathcal{N}\left(\mathbf{w}_0, \frac{\nu^2}{p} I_p\right) \quad (30)$$

The generating parameter \mathbf{w} is drawn once and kept fixed within a task, and drawn independently for different tasks. The values of \mathbf{x} and \mathbf{z} are drawn independently in all tasks and datasets (training, validation, target, test). In order to perform the calculations in the next section, we need the following results.

Lemma 1. Let X be a Gaussian $n \times p$ random matrix with independent rows, and each row has covariance equal to I_p , the $p \times p$ identity matrix. Then:

$$\mathbb{E} [X^T X] = n I_p \quad (31)$$

$$\mathbb{E} [(X^T X)^2] = n(n+p+1) I_p = n^2 \mu_2 I_p \quad (32)$$

$$\mathbb{E} [(X^T X)^3] = n(n^2 + p^2 + 3np + 3n + 3p + 4) I_p = n^3 \mu_3 I_p \quad (33)$$

$$\mathbb{E} [(X^T X)^4] = n(n^3 + p^3 + 6n^2 p + 6np^2 + \quad (34)$$

$$+ 6n^2 + 6p^2 + 17np + 21n + 21p + 20) I_p = n^4 \mu_4 I_p \quad (35)$$

$$\mathbb{E} [X^T X \text{Tr}(X^T X)] = (n^2 p + 2n) I_p = pn^2 \mu_{1,1} I_p \quad (36)$$

$$\mathbb{E} [(X^T X)^2 \text{Tr}(X^T X)] = n(n^2 p + np^2 + np + 4n + 4p + 4) I_p = pn^3 \mu_{2,1} I_p \quad (37)$$

$$\mathbb{E} [X^T X \text{Tr}((X^T X)^2)] = n(n^2 p + np^2 + np + 4n + 4p + 4) I_p = pn^3 \mu_{1,2} I_p \quad (38)$$

$$\mathbb{E} [(X^T X)^2 \text{Tr}((X^T X)^2)] = n(n^3 p + np^3 + 2n^2 p^2 + 2n^2 p + 2np^2 + \quad (39)$$

$$+ 8n^2 + 8p^2 + 21np + 20n + 20p + 20) I_p = pn^4 \mu_{2,2} I_p \quad (40)$$

where the last equality in each of these expressions defines the variables μ . Furthermore, for any $n \times n$ symmetric matrix C and any $p \times p$ symmetric matrix D , independent of X :

$$\mathbb{E} [X^T C X] = \text{Tr}(C) I_p \quad (41)$$

$$\mathbb{E} [X^T X D X^T X] = n(n+1) D + n \text{Tr}(D) I_p \quad (42)$$

Proof. The Lemma follows by direct computations of the above expectations, using Isserlis' theorem. Particularly, for higher order exponents, combinatorics plays a crucial role in counting products of different Gaussian variables in an effective way. □

Lemma 2. Let $X^{v(i)}$, $X^{t(i)}$ be Gaussian random matrices, of size respectively $n_v \times p$ and $n_t \times p$, with independent rows, and each row has covariance equal to I_p , the $p \times p$ identity matrix. Let p and n_t be large, both of order $o(\xi)$, where ξ is a large number. Then:

$$X^{v(i)} X^{v(i)T} = p I_{n_v} + o(\xi^{1/2}) \quad (43)$$

$$X^{v(i)} X^{t(i)T} X^{t(i)} X^{v(i)T} = pn_t I_{n_v} + o(\xi^{3/2}) \quad (44)$$

$$X^{v(i)} X^{t(i)T} X^{t(i)} X^{t(i)T} X^{t(i)} X^{v(i)T} = pn_t(n_t + p + 1) I_{n_v} + o(\xi^{5/2}) \quad (45)$$

Note that the order $o(\xi)$ applies to all elements of the matrix in each expression. For $i \neq j$

$$X^{v(i)} X^{v(j)T} = o(\xi^{1/2}) \quad (46)$$

$$X^{v(i)} X^{t(i)T} X^{t(i)} X^{v(j)T} = o(\xi^{3/2}) \quad (47)$$

$$X^{v(i)} X^{t(i)T} X^{t(i)} X^{t(j)T} X^{t(j)} X^{v(j)T} = o(\xi^{5/2}) \quad (48)$$

Furthermore, for any $p \times p$ symmetric matrix D independent of X , where the trace $\text{Tr}(D^2)$ is of order $o(\xi^\delta)$

$$X^{v(i)} D X^{v(i)T} = \text{Tr}(D) I_{n_v} + o(\xi^{\delta/2}) \quad (49)$$

$$X^{v(i)} X^{t(i)T} X^{t(i)} D X^{v(i)T} = \text{Tr}(D) n_t I_{n_v} + o(\xi^{1+\delta/2}) \quad (50)$$

$$X^{v(i)} X^{t(i)T} X^{t(i)} D X^{t(i)T} X^{t(i)} X^{v(i)T} = \text{Tr}(D) n_t (n_t + p + 1) I_{n_v} + o(\xi^{2+\delta/2}) \quad (51)$$

$$X^{v(i)} D X^{v(j)T} = o(\xi^{\delta/2}) \quad (52)$$

$$X^{v(i)} X^{t(i)T} X^{t(i)} D X^{v(j)T} = o(\xi^{1+\delta/2}) \quad (53)$$

$$X^{v(i)} X^{t(i)T} X^{t(i)} D X^{t(j)T} X^{t(j)} X^{v(j)T} = o(\xi^{2+\delta/2}) \quad (54)$$

Proof. The Lemma follows by direct computations of the expectations and variances of each term. \square

Lemma 3. Let X^v, X^t be Gaussian random matrices, of size respectively $n_v \times p$ and $n_t \times p$, with independent rows, and each row has covariance equal to I_p , the $p \times p$ identity matrix. Let n_v and n_t be large, both of order $o(\xi)$, where ξ is a large number. Then:

$$X^{vT} X^v = n_v I_p + o(\xi^{1/2}) \quad (55)$$

$$X^{tT} X^t X^{vT} X^v = n_t n_v I_p + o(\xi^{3/2}) \quad (56)$$

$$X^{tT} X^t X^{vT} X^v X^{tT} X^t = n_v n_t (n_t + p + 1) I_p + o(\xi^{5/2}) \quad (57)$$

Note that the order $o(\xi)$ applies to all elements of the matrix in each expression.

Proof. The Lemma follows by direct computations of the expectations and variances of each term. \square

7.3 PROOF OF THEOREMS 1 AND 2

We calculate the average test loss as a function of the hyperparameters $n_t, n_v, n_r, p, m, \alpha_t, \alpha_r, \sigma, \nu$. Using the expression in Eq.24 for the test output, we rewrite the test loss in Eq.27 as

$$\bar{\mathcal{L}}^{test} = \mathbb{E} \frac{1}{2n_s} |X^s (\mathbf{w}' - \boldsymbol{\theta}^*) + \mathbf{z}^s|^2 \quad (58)$$

We start by averaging this expression with respect to X^s, \mathbf{z}^s , noting that $\boldsymbol{\theta}^*$ does not depend on test data. We further average with respect to \mathbf{w}' , but note that $\boldsymbol{\theta}^*$ depends on test parameters, so we average only terms that do not depend on $\boldsymbol{\theta}^*$. Using Eq.31, the result is

$$\bar{\mathcal{L}}^{test} = \frac{\sigma^2}{2} + \frac{\nu^2}{2} + \frac{|\mathbf{w}_0|^2}{2} + \mathbb{E} \left[\frac{|\boldsymbol{\theta}^*|^2}{2} - (\mathbf{w}_0 + \delta \mathbf{w}')^T \boldsymbol{\theta}^* \right] \quad (59)$$

where we define $\delta \mathbf{w}' = \mathbf{w}' - \mathbf{w}_0$. The second term in the expectation is linear in $\boldsymbol{\theta}^*$ and can be averaged over X^r, \mathbf{z}^r , using Eq.25 and noting that $\boldsymbol{\omega}^*$ does not depend on target data. The result is

$$\mathbb{E}_{X^r} \mathbb{E}_{\mathbf{z}^r} \boldsymbol{\theta}^* = (1 - \alpha_r) \boldsymbol{\omega}^* + \alpha_r (\mathbf{w}_0 + \delta \mathbf{w}') \quad (60)$$

Using Eq.60 we average over \mathbf{w}' the second term in the expectation of Eq.59 and find

$$\bar{\mathcal{L}}^{test} = \frac{\sigma^2}{2} + \left(\frac{1}{2} - \alpha_r \right) \left(\nu^2 + |\mathbf{w}_0|^2 \right) - (1 - \alpha_r) \mathbf{w}_0^T \mathbb{E} \boldsymbol{\omega}^* + \mathbb{E} \frac{|\boldsymbol{\theta}^*|^2}{2} \quad (61)$$

We average the last term of this expression over $\mathbf{z}^r, \mathbf{w}'$, using Eq.25 and noting that ω^* does not depend on target data and test parameters. The result is

$$\mathbb{E}_{\mathbf{w}'} \mathbb{E}_{\mathbf{z}^r} |\theta^*|^2 = |\omega^*|^2 + \frac{\alpha_r^2}{n_r^2} (\omega^* - \mathbf{w}_0)^T \left(X^{rT} X^r \right)^2 (\omega^* - \mathbf{w}_0) - \quad (62)$$

$$- \frac{2\alpha_r}{n_r} X^{rT} X^r \omega^{*T} (\omega^* - \mathbf{w}_0) + \frac{\alpha_r^2 \sigma^2}{n_r^2} \text{Tr} \left[X^r X^{rT} \right] + \frac{\alpha_r^2 \nu^2}{n_r^2 p} \text{Tr} \left[\left(X^r X^{rT} \right)^2 \right] \quad (63)$$

We now average over X^r , again noting that ω^* does not depend on target data. Using Eqs.31, 32, we find

$$\mathbb{E}_{X^r} \mathbb{E}_{\mathbf{w}'} \mathbb{E}_{\mathbf{z}^r} |\theta^*|^2 = |\omega^*|^2 + \alpha_r^2 \left(1 + \frac{p+1}{n_r} \right) \left(\nu^2 + |\omega^* - \mathbf{w}_0|^2 \right) - 2\alpha_r \omega^{*T} (\omega^* - \mathbf{w}_0) + \frac{\alpha_r^2 \sigma^2 p}{n_r} \quad (64)$$

We can now rewrite the average test loss 61 as

$$\bar{\mathcal{L}}^{test} = \frac{\sigma^2}{2} \left(1 + \frac{\alpha_r^2 p}{n_r} \right) + \frac{1}{2} \left[(1 - \alpha_r)^2 + \alpha_r^2 \frac{p+1}{n_r} \right] \left(\nu^2 + \mathbb{E} |\omega^* - \mathbf{w}_0|^2 \right) \quad (65)$$

In order to average the last term, we need an expression for ω^* . We note that the loss in Eq.20 is quadratic in ω , therefore the solution of Eq.22 can be found using standard linear algebra. In particular, the loss in Eq.20 can be rewritten as

$$\mathcal{L}^{meta} = \frac{1}{2n_v m} |\gamma - B\omega|^2 \quad (66)$$

where γ is a vector of shape $n_v m \times 1$, and B is a matrix of shape $n_v m \times p$. The vector γ is a stack of m vectors

$$\gamma = \begin{pmatrix} X^{v(1)} \left(I_p - \frac{\alpha_t}{n_t} X^{t(1)T} X^{t(1)} \right) \mathbf{w}^{(1)} - \frac{\alpha_t}{n_t} X^{v(1)} X^{t(1)T} \mathbf{z}^{t(1)} + \mathbf{z}^{v(1)} \\ \vdots \\ X^{v(m)} \left(I_p - \frac{\alpha_t}{n_t} X^{t(m)T} X^{t(m)} \right) \mathbf{w}^{(m)} - \frac{\alpha_t}{n_t} X^{v(m)} X^{t(m)T} \mathbf{z}^{t(m)} + \mathbf{z}^{v(m)} \end{pmatrix} \quad (67)$$

Similarly, the matrix B is a stack of m matrices

$$B = \begin{pmatrix} X^{v(1)} \left(I_p - \frac{\alpha_t}{n_t} X^{t(1)T} X^{t(1)} \right) \\ \vdots \\ X^{v(m)} \left(I_p - \frac{\alpha_t}{n_t} X^{t(m)T} X^{t(m)} \right) \end{pmatrix} \quad (68)$$

We denote by I_p the $p \times p$ identity matrix. The expression for ω that minimizes Eq.66 depends on whether the problem is overparameterized ($p > n_v m$) or underparameterized ($p < n_v m$), therefore we distinguish these two cases in the following sections.

7.3.1 OVERPARAMETERIZED CASE (THEOREM 1)

In the overparameterized case ($p > n_v m$), under the assumption that the inverse of BB^T exists, the value of ω that minimizes Eq.66 is equal to

$$\omega^* = B^T (BB^T)^{-1} \gamma + \left[I_p - B^T (BB^T)^{-1} B \right] \omega_0 \quad (69)$$

The vector ω_0 is interpreted as the initial condition of the parameter optimization of the outer loop, when optimized by gradient descent. Note that the matrix B does not depend on $\mathbf{w}, \mathbf{z}^t, \mathbf{z}^v$, and $\mathbb{E}_{\mathbf{w}} \mathbb{E}_{\mathbf{z}^t} \mathbb{E}_{\mathbf{z}^v} \gamma = B\mathbf{w}_0$. We denote by $\delta\gamma$ the deviation from the average, and we have

$$\omega^* - \mathbf{w}_0 = B^T (BB^T)^{-1} \delta\gamma + \left[I_p - B^T (BB^T)^{-1} B \right] (\omega_0 - \mathbf{w}_0) \quad (70)$$

We square this expression and average over $\mathbf{w}, \mathbf{z}^t, \mathbf{z}^v$. We use the cyclic property of the trace and the fact that $B^T (BB^T)^{-1} B$ is a projection. The result is

$$|\omega^* - \mathbf{w}_0|^2 = \text{Tr} \left[\Gamma (BB^T)^{-1} \right] + (\omega_0 - \mathbf{w}_0)^T \left[I_p - B^T (BB^T)^{-1} B \right] (\omega_0 - \mathbf{w}_0) \quad (71)$$

The matrix Γ is defined as

$$\Gamma = \mathbb{E}_{\mathbf{w}} \mathbb{E}_{\mathbf{z}^t} \mathbb{E}_{\mathbf{z}^v} \delta \gamma \delta \gamma^T = \begin{pmatrix} \Gamma^{(1)} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Gamma^{(m)} \end{pmatrix} \quad (72)$$

Where matrix blocks are given by the following expression

$$\Gamma^{(i)} = \frac{\nu^2}{p} X^{v(i)} \left(I_p - \frac{\alpha_t}{n_t} X^{t(i)T} X^{t(i)} \right)^2 X^{v(i)T} + \sigma^2 \left(I_{n_v} + \frac{\alpha_t^2}{n_t^2} X^{v(i)} X^{t(i)T} X^{t(i)} X^{v(i)T} \right) \quad (73)$$

It is convenient to rewrite the scalar product of Eq.71 in terms of the trace of outer products

$$|\boldsymbol{\omega}^* - \mathbf{w}_0|^2 = \text{Tr} \left[(BB^T)^{-1} \left(\Gamma - B(\boldsymbol{\omega}_0 - \mathbf{w}_0)(\boldsymbol{\omega}_0 - \mathbf{w}_0)^T B^T \right) \right] + |\boldsymbol{\omega}_0 - \mathbf{w}_0|^2 \quad (74)$$

In order to calculate $\mathbb{E} |\boldsymbol{\omega}^* - \mathbf{w}_0|^2$ in Eq.65 we need to average this expression over training and validation data. These averages are hard to compute since they involve nonlinear functions of the data. However, we can approximate these terms by assuming that p and n_t are large, both of order $o(\xi)$, where ξ is a large number. Furthermore, we assume that $|\boldsymbol{\omega}_0 - \mathbf{w}_0|$ is of order $o(\xi^{-1/4})$. Using Lemma 2, together with the expressions of B (Eq.68) and Γ (Eqs.72,73), we can prove that

$$\frac{1}{p} BB^T = \left[(1 - \alpha_t)^2 + \alpha_t^2 \frac{p+1}{n_t} \right] I_{n_v m} + o(\xi^{-1/2}) \quad (75)$$

$$\Gamma = \left\{ \nu^2 \left[(1 - \alpha_t)^2 + \alpha_t^2 \frac{p+1}{n_t} \right] + \sigma^2 \left(1 + \frac{\alpha_t^2 p}{n_t} \right) \right\} I_{n_v m} + o(\xi^{-1/2}) \quad (76)$$

$$B(\boldsymbol{\omega}_0 - \mathbf{w}_0)(\boldsymbol{\omega}_0 - \mathbf{w}_0)^T B^T = |\boldsymbol{\omega}_0 - \mathbf{w}_0|^2 \left[(1 - \alpha_t)^2 + \alpha_t^2 \frac{p+1}{n_t} \right] I_{n_v m} + o(\xi^{-1/2}) \quad (77)$$

Using Eq.75 and Taylor expansion, the inverse $(BB^T)^{-1}$ is equal to

$$(BB^T)^{-1} = \frac{1}{p} \left[(1 - \alpha_t)^2 + \alpha_t^2 \frac{p+1}{n_t} \right]^{-1} I_{n_v m} + o(\xi^{-3/2}), \quad (78)$$

Substituting the three expressions above in Eq.74, and ignoring terms of lower order, we find

$$\mathbb{E} |\boldsymbol{\omega}^* - \mathbf{w}_0|^2 = \left(1 - \frac{n_v m}{p} \right) |\boldsymbol{\omega}_0 - \mathbf{w}_0|^2 + \frac{n_v m}{p} \left[\nu^2 + \sigma^2 \frac{1 + \frac{\alpha_t^2 p}{n_t}}{(1 - \alpha_t)^2 + \alpha_t^2 \frac{p+1}{n_t}} \right] + o(\xi^{-3/2}) \quad (79)$$

Substituting this expression into in Eq.65, we find the value of average test loss

$$\begin{aligned} \bar{\mathcal{L}}^{test} &= \frac{\sigma^2}{2} \left(1 + \frac{\alpha_r^2 p}{n_r} \right) + \\ &+ h^r \left[\frac{\nu^2}{2} \left(1 + \frac{n_v m}{p} \right) + \frac{1}{2} \left(1 - \frac{n_v m}{p} \right) |\boldsymbol{\omega}_0 - \mathbf{w}_0|^2 + \frac{\sigma^2 n_v m}{2p} \frac{1 + \frac{\alpha_t^2 p}{n_t}}{h^t} \right] + o(\xi^{-3/2}) \end{aligned} \quad (80)$$

(81)

where we define the following expressions

$$h^t = (1 - \alpha_t)^2 + \alpha_t^2 \frac{p+1}{n_t} \quad \text{and} \quad h^r = (1 - \alpha_r)^2 + \alpha_r^2 \frac{p+1}{n_r} \quad (82)$$

7.3.2 UNDERPARAMETERIZED CASE (THEOREM 2)

In the underparameterized case ($p < n_v m$), under the assumption that the inverse of $B^T B$ exists, the value of $\boldsymbol{\omega}$ that minimizes Eq.66 is equal to

$$\boldsymbol{\omega}^* = (B^T B)^{-1} B^T \gamma \quad (83)$$

Note that the matrix B does not depend on $\mathbf{w}, \mathbf{z}^t, \mathbf{z}^v$, and $\mathbb{E}_{\mathbf{w}} \mathbb{E}_{\mathbf{z}^t} \mathbb{E}_{\mathbf{z}^v} \gamma = B\mathbf{w}_0$. We denote by $\delta\gamma$ the deviation from the average, and we have

$$|\omega^* - \mathbf{w}_0|^2 = \text{Tr} \left[(B^T B)^{-1} B^T \delta\gamma \delta\gamma^T B (B^T B)^{-1} \right] \quad (84)$$

We need to average this expression in order to calculate $\mathbb{E} |\omega^* - \mathbf{w}_0|^2$ in Eq.65. We start by averaging $\delta\gamma \delta\gamma^T$ over $\mathbf{w}, \mathbf{z}^t, \mathbf{z}^v$, since B does not depend on those variables. Note that $\mathbf{w}, \mathbf{z}^t, \mathbf{z}^v$ are independent on each other and across tasks. As in previous section, we denote by Γ the result of this operation, given by Eqs.72, 73. Finally, we need to average over the training and validation data

$$\mathbb{E} |\omega^* - \mathbf{w}_0|^2 = \mathbb{E}_{X^t} \mathbb{E}_{X^v} \text{Tr} \left[(B^T B)^{-1} B^T \Gamma B (B^T B)^{-1} \right] \quad (85)$$

It is hard to average this expression because it includes nonlinear functions of the data. However, we can approximate these terms by assuming that either m or ξ (or both) is a large number, where ξ is defined by assuming that both n_t and n_v are of order $o(\xi)$. Using Lemma 3, together with the expression of B (Eq.68), and noting that each factor in Eq.85 has a sum over m independent terms, we can prove that

$$\frac{1}{n_v m} B^T B = (1 - 2\alpha_t + \alpha_t^2 \mu_2) I_p + o\left((m\xi)^{-1/2}\right) \quad (86)$$

The expression for μ_2 is given in Eq.32. Using this result and a Taylor expansion, the inverse is equal to

$$n_v m (B^T B)^{-1} = (1 - 2\alpha_t + \alpha_t^2 \mu_2)^{-1} I_p + o\left((m\xi)^{-1/2}\right) \quad (87)$$

Similarly, the term $B^T \Gamma B$ is equal to its average plus a term of smaller order

$$\frac{1}{n_v m} B^T \Gamma B = \frac{1}{n_v m} \mathbb{E} (B^T \Gamma B) + o\left((m\xi)^{-1/2}\right) \quad (88)$$

We substitute these expressions in Eq.85 and neglect lower orders. Here we show how to calculate explicitly the expectation of $B^T \Gamma B$. For ease of notation, we define the matrix $A^{t(i)} = I - \frac{\alpha_t}{n_t} X^{t(i)T} X^{t(i)}$. Using the expressions of B (Eq.68) and Γ (Eqs.72,73), the expression for $B^T \Gamma B$ is given by

$$\begin{aligned} B^T \Gamma B &= \sigma^2 \sum_{i=1}^m A^{t(i)T} X^{v(i)T} X^{v(i)} A^{t(i)} + \frac{\nu^2}{p} \sum_{i=1}^m \left(A^{t(i)T} X^{v(i)T} X^{v(i)} A^{t(i)} \right)^2 + \\ &+ \frac{\alpha_t^2 \sigma^2}{n_t^2} \sum_{i=1}^m A^{t(i)T} X^{v(i)T} X^{v(i)} X^{t(i)T} X^{t(i)} X^{v(i)T} X^{v(i)} A^{t(i)} \end{aligned} \quad (89)$$

We use Eqs.31, 32 to calculate the average of the first term in Eq.89

$$\mathbb{E}_{X^t} \mathbb{E}_{X^v} \sum_{i=1}^m A^{t(i)T} X^{v(i)T} X^{v(i)} A^{t(i)} = n_v m (1 - 2\alpha_t + \alpha_t^2 \mu_2) I_p \quad (90)$$

We use Eqs.31, 32, 33, 41, 36, 37, 38, 39 to calculate the average of the second term

$$\mathbb{E}_{X^t} \mathbb{E}_{X^v} \sum_{i=1}^m \left(A^{t(i)T} X^{v(i)T} X^{v(i)} A^{t(i)} \right)^2 = \mathbb{E}_{X^t} \sum_{i=1}^m \left[n_v (n_v + 1) A^{t(i)4} + n_v A^{t(i)2} \text{Tr} \left(A^{t(i)2} \right) \right] = \quad (91)$$

$$\begin{aligned} &= mn_v (n_v + 1) (1 - 4\alpha_t + 6\alpha_t^2 \mu_2 - 4\alpha_t^3 \mu_3 + \alpha_t^4 \mu_4) I_p + \\ &+ mn_v p (1 - 4\alpha_t + 2\alpha_t^2 \mu_2 + 4\alpha_t^2 \mu_{1,1} - 4\alpha_t^3 \mu_{2,1} + \alpha_t^4 \mu_{2,2}) I_p \end{aligned} \quad (92)$$

Finally, we compute the average of the third term, using Eqs.31, 32, 33, 34, 41, 36, 37

$$\mathbb{E}_{X^t} \mathbb{E}_{X^v} \sum_{i=1}^m A^{t(i)T} X^{v(i)T} X^{v(i)} X^{t(i)T} X^{t(i)} X^{v(i)T} X^{v(i)} A^{t(i)} = \quad (93)$$

$$= \mathbb{E}_{X^t} \sum_{i=1}^m \left[n_v (n_v + 1) A^{t(i)T} X^{t(i)T} X^{t(i)} A^{t(i)} + n_v A^{t(i)T} A^{t(i)} \text{Tr} \left(X^{t(i)T} X^{t(i)} \right) \right] = \quad (94)$$

$$= mn_v (n_v + 1) n_t (1 - 2\alpha_t \mu_2 + \alpha_t^2 \mu_3) I_p + mn_v n_t p (1 - 2\alpha_t \mu_{1,1} + \alpha_t^2 \mu_{2,1}) I_p \quad (95)$$

Putting everything together in Eq.85, and applying the trace operator, we find the following expression for the meta-parameter variance

$$\begin{aligned}\mathbb{E}|\boldsymbol{\omega}^* - \mathbf{w}_0|^2 &= \frac{p}{n_v m} (1 - 2\alpha_t + \alpha_t^2 \mu_2)^{-2} \left\{ \sigma^2 (1 - 2\alpha_t + \alpha_t^2 \mu_2) + \right. \\ &+ \frac{\alpha_t^2 \sigma^2}{n_t} [(n_v + 1) (1 - 2\alpha_t \mu_2 + \alpha_t^2 \mu_3) + p (1 - 2\alpha_t \mu_{1,1} + \alpha_t^2 \mu_{2,1})] \\ &+ \frac{\nu^2}{p} \left[(n_v + 1) (1 - 4\alpha_t + 6\alpha_t^2 \mu_2 - 4\alpha_t^3 \mu_3 + \alpha_t^4 \mu^4) + \right. \\ &\left. \left. + p (1 - 4\alpha_t + 2\alpha_t^2 \mu_2 + 4\alpha_t^2 \mu_{1,1} - 4\alpha_t^3 \mu_{2,1} + \alpha_t^4 \mu_{2,2}) \right] \right\} + o((m\xi)^{-3/2})\end{aligned}\quad (96)$$

We rewrite this expression as

$$\begin{aligned}\mathbb{E}|\boldsymbol{\omega}^* - \mathbf{w}_0|^2 &= \frac{p}{h^t n_v m} \left\{ \sigma^2 \left[h^t + \frac{\alpha_t^2}{n_t} [(n_v + 1) g_1 + p g_2] \right] + \frac{\nu^2}{p} [(n_v + 1) g_3 + p g_3] \right\} + \\ &+ o((m\xi)^{-3/2})\end{aligned}\quad (97)$$

where we defined the following expressions for g_i

$$g_1 = 1 - 2\alpha_t \mu_2 + \alpha_t^2 \mu_3 \quad (98)$$

$$g_2 = 1 - 2\alpha_t \mu_{1,1} + \alpha_t^2 \mu_{2,1} \quad (99)$$

$$g_3 = 1 - 4\alpha_t + 6\alpha_t^2 \mu_2 - 4\alpha_t^3 \mu_3 + \alpha_t^4 \mu^4 \quad (100)$$

$$g_4 = 1 - 4\alpha_t + 2\alpha_t^2 \mu_2 + 4\alpha_t^2 \mu_{1,1} - 4\alpha_t^3 \mu_{2,1} + \alpha_t^4 \mu_{2,2} \quad (101)$$

and μ_i are equal to

$$\mu_2 = \frac{1}{n_t} (n_t + p + 1) \quad (102)$$

$$\mu_3 = \frac{1}{n_t^2} (n_t^2 + p^2 + 3n_t p + 3n_t + 3p + 4) \quad (103)$$

$$\mu_4 = \frac{1}{n_t^3} (n_t^3 + p^3 + 6n_t^2 p + 6n_t p^2 + 6n_t^2 + 6p^2 + 17n_t p + 21n_t + 21p + 20) \quad (104)$$

$$\mu_{1,1} = \frac{1}{n_t^2 p} (n_t^2 p + 2n_t) \quad (105)$$

$$\mu_{2,1} = \frac{1}{n_t^2 p} (n_t^2 p + n_t p^2 + n_t p + 4n_t + 4p + 4) \quad (106)$$

$$\mu_{2,2} = \frac{1}{n_t^3 p} (n_t^3 p + n_t p^3 + 2n_t^2 p^2 + 2n_t^2 p + 2n_t p^2 + 8n_t^2 + 8p^2 + 21n_t p + 20n_t + 20p + 20) \quad (107)$$

Substituting this expression back into Eq.65 returns the final expression for the average test loss, equal to

$$\begin{aligned}\bar{\mathcal{L}}^{test} &= \frac{\sigma^2}{2} \left(1 + \frac{\alpha_r^2 p}{n_r} \right) + \frac{h^r \nu^2}{2} + \\ &+ \frac{h^r}{2h^t^2} \frac{p}{n_v m} \left\{ \sigma^2 \left[h^t + \frac{\alpha_t^2}{n_t} [(n_v + 1) g_1 + p g_2] \right] + \frac{\nu^2}{p} [(n_v + 1) g_3 + p g_4] \right\} + o((m\xi)^{-3/2})\end{aligned}\quad (108)$$

7.4 PROOF OF THEOREM 3

In this section, we release some assumption on the distributions of data and parameters. In particular, we do not assume a specific distribution for input data vectors \mathbf{x} and generating parameter vector

\mathbf{w} , besides that different data vectors are independent, and so are data and parameters for different tasks. We further assume that those vectors have zero mean, and denote their covariance as

$$\Sigma = \mathbb{E} \mathbf{x} \mathbf{x}^T \quad (109)$$

$$\Sigma_w = \mathbb{E} \mathbf{w} \mathbf{w}^T \quad (110)$$

We will also use the following matrix, including fourth order moments

$$F = \mathbb{E} (\mathbf{x}^T \Sigma \mathbf{x}) \mathbf{x} \mathbf{x}^T \quad (111)$$

We do not make any assumption about the distribution of \mathbf{x} , but we note that, if \mathbf{x} is Gaussian, then $F = 2\Sigma^3 + \Sigma \text{Tr}(\Sigma^2)$. We keep the assumption that the output noise is Gaussian and independent for different data points and tasks, with variance σ^2 . Using the same notation as in previous sections, we will also use the following expressions (for any $p \times p$ matrix A)

$$\mathbb{E} [X^T X] = n\Sigma \quad (112)$$

$$\mathbb{E} \text{Tr} [\Sigma X^T X A X^T X] = \text{Tr} \{A [n^2 \Sigma^3 + n(F - \Sigma^3)]\} \quad (113)$$

We proceed to derive the same formula under these less restrictive assumptions, in the overparameterized case only, following is the same derivation of section 7.3. We further assume $\omega_0 = 0$, $\mathbf{w}_0 = 0$. Again we start from the expression in Eq.24 for the test output, and we rewrite the test loss in Eq.27 as

$$\bar{\mathcal{L}}^{test} = \mathbb{E} \frac{1}{2n_s} |X^s (\mathbf{w}' - \boldsymbol{\theta}^*) + \mathbf{z}^s|^2 \quad (114)$$

We average this expression with respect to X^s, \mathbf{z}^s , noting that $\boldsymbol{\theta}^*$ does not depend on test data. We further average with respect to \mathbf{w}' , but note that $\boldsymbol{\theta}^*$ depends on test parameters, so we average only terms that do not depend on $\boldsymbol{\theta}^*$. Using Eq.112, the result is

$$\bar{\mathcal{L}}^{test} = \frac{\sigma^2}{2} + \frac{1}{2} \text{Tr}(\Sigma \Sigma_w) + \mathbb{E} \left[\frac{1}{2} \boldsymbol{\theta}^{*T} \Sigma \boldsymbol{\theta}^* - \mathbf{w}'^T \Sigma \boldsymbol{\theta}^* \right] \quad (115)$$

The second term in the expectation is linear in $\boldsymbol{\theta}^*$ and can be averaged over X^r, \mathbf{z}^r , using Eq.25 and noting that ω^* does not depend on target data. The result is

$$\mathbb{E}_{X^r} \mathbb{E}_{\mathbf{z}^r} \boldsymbol{\theta}^* = (I - \alpha_r \Sigma) \omega^* + \alpha_r \Sigma \mathbf{w}' \quad (116)$$

Furthermore, we show below (Eq.128) that the following average holds

$$\mathbb{E}_{\mathbf{w}} \mathbb{E}_{\mathbf{z}^t} \mathbb{E}_{\mathbf{z}^v} \omega^* = 0 \quad (117)$$

Combining Eqs.116, 117, we can calculate the second term in the expectation of Eq.115 and find

$$\bar{\mathcal{L}}^{test} = \frac{\sigma^2}{2} + \frac{1}{2} \text{Tr}(\Sigma \Sigma_w) - \alpha_r \text{Tr}(\Sigma^2 \Sigma_w) + \mathbb{E} \frac{1}{2} \boldsymbol{\theta}^{*T} \Sigma \boldsymbol{\theta}^* \quad (118)$$

We start by averaging the third term of this expression over $\mathbf{z}^r, \mathbf{w}'$, using Eq.25 and noting that ω^* does not depend on target data and test parameters. The result is

$$\mathbb{E}_{\mathbf{w}'} \mathbb{E}_{\mathbf{z}^r} \boldsymbol{\theta}^{*T} \Sigma \boldsymbol{\theta}^* = \text{Tr} \left[\Sigma \left(I - \frac{\alpha_r}{n_r} X^{rT} X^r \right) \omega^* \omega^{*T} \left(I - \frac{\alpha_r}{n_r} X^{rT} X^r \right) \right] + \quad (119)$$

$$+ \frac{\alpha_r^2 \sigma^2}{n_r^2} \text{Tr} [X^r \Sigma X^{rT}] + \frac{\alpha_r^2}{n_r^2} \text{Tr} [\Sigma X^{rT} X^r \Sigma_w X^{rT} X^r] \quad (120)$$

We now average over X^r , again noting that ω^* does not depend on target data. Using Eqs.112, 113, we find

$$\mathbb{E}_{X^r} \mathbb{E}_{\mathbf{w}'} \mathbb{E}_{\mathbf{z}^r} \boldsymbol{\theta}^{*T} \Sigma \boldsymbol{\theta}^* = \text{Tr} \left\{ \omega^* \omega^{*T} \left[\Sigma (I - \alpha_r \Sigma)^2 + \frac{\alpha_r^2}{n_r} (F - \Sigma^3) \right] \right\} + \quad (121)$$

$$+ \frac{\alpha_r^2 \sigma^2}{n_r} \text{Tr}(\Sigma^2) + \alpha_r^2 \text{Tr} \left\{ \Sigma_w \left[\Sigma^3 + \frac{1}{n_r} (F - \Sigma^3) \right] \right\} \quad (122)$$

We can now rewrite the average test loss in Eq.118 as

$$\bar{\mathcal{L}}^{test} = \frac{\sigma^2}{2} \left[1 + \frac{\alpha_r^2}{n_r} \text{Tr}(\Sigma^2) \right] + \frac{1}{2} \text{Tr} \left[\left(\Sigma_w + \mathbb{E} \omega^* \omega^{*T} \right) H^r \right] \quad (123)$$

where we define the following matrix

$$H^r = \left[\Sigma (I - \alpha_r \Sigma)^2 + \frac{\alpha_r^2}{n_r} (F - \Sigma^3) \right] \quad (124)$$

In order to average the last term, we need an expression for ω^* . We note that the loss in Eq.20 is quadratic in ω , therefore the solution in Eq.22 can be found using standard linear algebra. In particular, the loss in Eq.20 can be rewritten as

$$\mathcal{L}^{meta} = \frac{1}{2n_v m} |\gamma - B\omega|^2 \quad (125)$$

where γ is a vector of shape $n_v m \times 1$, and B is a matrix of shape $n_v m \times p$. The vector γ is a stack of m vectors

$$\gamma = \begin{pmatrix} X^{v(1)} \left(I - \frac{\alpha_t}{n_t} X^{t(1)T} X^{t(1)} \right) \mathbf{w}^{(1)} - \frac{\alpha_t}{n_t} X^{v(1)} X^{t(1)T} \mathbf{z}^{t(1)} + \mathbf{z}^{v(1)} \\ \vdots \\ X^{v(m)} \left(I - \frac{\alpha_t}{n_t} X^{t(m)T} X^{t(m)} \right) \mathbf{w}^{(m)} - \frac{\alpha_t}{n_t} X^{v(m)} X^{t(m)T} \mathbf{z}^{t(m)} + \mathbf{z}^{v(m)} \end{pmatrix} \quad (126)$$

Similarly, the matrix B is a stack of m matrices

$$B = \begin{pmatrix} X^{v(1)} \left(I - \frac{\alpha_t}{n_t} X^{t(1)T} X^{t(1)} \right) \\ \vdots \\ X^{v(m)} \left(I - \frac{\alpha_t}{n_t} X^{t(m)T} X^{t(m)} \right) \end{pmatrix} \quad (127)$$

In the overparameterized case ($p > n_v m$), under the assumption that the inverse of BB^T exists, the value of ω that minimizes Eq.125, and that also has minimum norm, is equal to

$$\omega^* = B^T (BB^T)^{-1} \gamma \quad (128)$$

Note that the matrix B does not depend on \mathbf{w} , \mathbf{z}^t , \mathbf{z}^v , and $\mathbb{E}_{\mathbf{w}} \mathbb{E}_{\mathbf{z}^t} \mathbb{E}_{\mathbf{z}^v} \gamma = 0$, therefore Eq.117 holds. In order to finish calculating Eq.123, we need to average the following term

$$\text{Tr} \left(H^r \omega^* \omega^{*T} \right) = \text{Tr} \left[(BB^T)^{-1} \gamma \gamma^T (BB^T)^{-1} (B H^r B^T) \right] \quad (129)$$

where we used the cyclic property of the trace. We start by averaging $\gamma \gamma^T$ over \mathbf{w} , \mathbf{z}^t , \mathbf{z}^v , since B does not depend on those variables. Note that \mathbf{w} , \mathbf{z}^t , \mathbf{z}^v are independent on each other and across tasks. We denote by Γ the result of this operation, which is equal to a block diagonal matrix

$$\Gamma = \mathbb{E}_{\mathbf{w}} \mathbb{E}_{\mathbf{z}^t} \mathbb{E}_{\mathbf{z}^v} \gamma \gamma^T = \begin{pmatrix} \Gamma^{(1)} & 0 & 0 \\ 0 & \ddots & 0 \\ 0 & 0 & \Gamma^{(m)} \end{pmatrix} \quad (130)$$

Where matrix blocks are given by the following expression

$$\Gamma^{(i)} = X^{v(i)} \left(I - \frac{\alpha_t}{n_t} X^{t(i)T} X^{t(i)} \right) \Sigma_w \left(I - \frac{\alpha_t}{n_t} X^{t(i)T} X^{t(i)} \right) X^{v(i)T} + \quad (131)$$

$$+ \sigma^2 \left(I_{n_v} + \frac{\alpha_t^2}{n_t^2} X^{v(i)} X^{t(i)T} X^{t(i)} X^{v(i)T} \right) \quad (132)$$

Finally, we need to average over the training and validation data

$$\mathbb{E} \text{Tr} \left(H^r \omega^* \omega^{*T} \right) = \mathbb{E}_{X^t} \mathbb{E}_{X^v} \text{Tr} \left[(BB^T)^{-1} \Gamma (BB^T)^{-1} (B H^r B^T) \right] \quad (133)$$

These averages are hard to compute since they involve nonlinear functions of the data. However, we can approximate these terms by assuming that p and n_t are large, both of order $o(\xi)$, where ξ is a large number. Furthermore, we assume that $\text{Tr}(\Sigma_w^2)$ is of order $o(\xi^{-1})$, and that the variances of matrix products of the rescaled inputs \mathbf{x}/\sqrt{p} , up to sixth order, are all of order $o(\xi^{-1})$, in particular

$$\text{Var}\left(\frac{1}{p}X^{v(i)}X^{v(j)T}\right) = o(\xi^{-1}) \quad (134)$$

$$\text{Var}\left(\frac{1}{p^2}X^{v(i)}X^{t(i)T}X^{t(i)}X^{v(j)T}\right) = o(\xi^{-1}) \quad (135)$$

$$\text{Var}\left(\frac{1}{p^3}X^{v(i)}X^{t(i)T}X^{t(i)}X^{t(j)T}X^{t(j)}X^{v(j)T}\right) = o(\xi^{-1}) \quad (136)$$

Then, using Eqs.112, 113 and the expressions of B (Eq.127) and Γ (Eqs.130,131), we can prove that

$$BB^T = \text{Tr}(H^t)I_{n_v m} + o(\xi^{1/2}) \quad (137)$$

$$\Gamma = \left\{ \text{Tr}(\Sigma_w H^t) + \sigma^2 \left[1 + \frac{\alpha_t^2}{n_t} \text{Tr}(\Sigma^2) \right] \right\} I_{n_v m} + o(\xi^{1/2}) \quad (138)$$

$$BH^r B^T = \text{Tr}(H^r H^t) I_{n_v m} + o(\xi^{1/2}) \quad (139)$$

where, similar to Eq.124, we define

$$H^t = \left[\Sigma(I - \alpha_t \Sigma)^2 + \frac{\alpha_t^2}{n_t} (F - \Sigma^3) \right] \quad (140)$$

Note that all these terms are of order $o(\xi)$. The inverse of BB^T can be found by a Taylor expansion

$$(BB^T)^{-1} = \text{Tr}(H^t)^{-1} I_{n_v m} + o(\xi^{-3/2}) \quad (141)$$

Substituting these expressions in Eq.133, we find

$$\mathbb{E} \text{Tr}(H^r \omega^* \omega^{*T}) = n_v m \frac{\text{Tr}(H^r H^t) \left\{ \text{Tr}(\Sigma_w H^t) + \sigma^2 \left[1 + \frac{\alpha_t^2}{n_t} \text{Tr}(\Sigma^2) \right] \right\}}{\text{Tr}(H^t)^2} + o(\xi^{-3/2}) \quad (142)$$

Substituting this expression into in Eq.123, we find the value of average test loss

$$\bar{\mathcal{L}}^{test} = \frac{1}{2} \text{Tr}(\Sigma_w H^r) + \frac{\sigma^2}{2} \left[1 + \frac{\alpha_r^2}{n_r} \text{Tr}(\Sigma^2) \right] + \quad (143)$$

$$+ \frac{1}{2} n_v m \frac{\text{Tr}(H^r H^t) \left\{ \text{Tr}(\Sigma_w H^t) + \sigma^2 \left[1 + \frac{\alpha_t^2}{n_t} \text{Tr}(\Sigma^2) \right] \right\}}{\text{Tr}(H^t)^2} + o(\xi^{-3/2}) \quad (144)$$