Studies on Neural Networks: Information Propagation at Initialisation and Robustness to Adversarial Examples



Giuseppe Ughi

Wolfson College

University of Oxford

A thesis submitted for the degree of

Doctor of Philosophy

Hillary 2022

Alla mia Mamma,
un costante esempio a cui devo tutto.

Acknowledgements

I would like to first express my gratitude to my supervisor, Jared Tanner, who has supported me throughout these four years of research, allowing me to grow from considering deep learning like sorcery to actually enjoying the study of its mathematical structure.

I am also extremely thankful to New Rock Capital Managment/RE5Q. Especially, I need to thank Richard Hodgson and Seth Rogers who both acted as industrial supervisors for my project. Their willingness to allow me to primarily focus on mathematical questions has been invaluable to the final output of my research. Thanks to the exciting environment that they nurtured within the company, I will remember all the visits with great fondness.

This thesis was made possible thanks to the Industrially Focused Mathematical Modelling course directors, Chris Breward and Colin Please. I hope that their effort in making us better mathematicians has not fallen short with me.

A special thanks goes to Mary P. Sanford, who made all this experience worthwhile. I'm looking forward to seeing what the future holds for us. A fellow acknowledgement also to the "the good, the ugly, il cattivo" gang at 24 Wharf House. Whilst the world was crumbling under Covid, thanks to you I felt rarely as wholesome staying at home.

These four years flew by thanks to all the friends who always supported me and encouraged me to move forward. In this context I want to thank all the colleagues/friends in InFoMM and at New Rock. Further thanks to old friends that never get old, Eduardo C., Edoardo P., Giovanni G, Tommaso F., Luigi R., Tito B., Pierre R., Giody, Valerio C., Giacomo M., and all of the Wolfson groups from whom I cannot thank any individual enough for the ride.

And last but by no means least, I want to thank my family for their support in all the activities that have led me here. I could not have asked for a better surrounding in which to grow.

Abstract

Over the last decade, the academic and industrial communities have become increasingly involved in the field of Deep Learning, leading these algorithms to become the drivers of the AI revolution that has allowed "machines" to attain increasingly extraordinary results. However, despite the impressive results achieved, there is still room for these algorithms to improve, and the use of mathematical techniques is central to this advancement. In this thesis we focus on two different themes in this space: the initialisation and the robustness of neural networks.

Currently the training of these algorithms is becoming an always more expensive procedure, not only financially, but also ecologically. To speed-up this process and therefore alleviate the above problems, one can focus on trying to more optimally initialise the neural networks. In this thesis we focus on defining this optimality with measures taken from information theory. Specifically, we introduce a lower-bound on the mutual information of the neural networks and find that there is an initialisation procedure that allows signals to propagate through the neural networks with less loss of information.

Finally, neural networks are vulnerable to adversarial attacks, meaning that regular inputs with small perturbations, although appearing to be insignificant to humans, can cause undesired model behaviours. In particular, we focus on identifying the methods that generate these adversarial perturbations, considering the neural networks as black-boxes. To do this, we consider the state-of-the-art methods and compare them for the first time in the same exact settings. We also introduce a further method to ensure that all the most common derivative-free-optimisation techniques have been tested on this problem. Thanks to our experiments, we are able to show how there is a class of methods that is particularly effective at generating these adversarial perturbations, but also how the various methods are affected differently by the setting in which they are deployed, therefore indicating that there is not a one-size-fits-all solution.

Contents

1 Introduction **1**

1.1 Neural Networks 2

1.1.1 Initialisation of Neural Networks 4

1.1.2 Adversarial Examples 6

1.2 Research Objectives and Contributions 7

1.3 Thesis Structure 8

1.4 Collaboration with New Rock Management 10

2 A Mean-Field Approximation of the Mutual Information of Neural Networks at Initialisation **12**

2.1 Introduction to Random Neural Networks 13

2.1.1 Geometric Flow 13

2.1.2 Gradient Flow 15

2.1.3 Mutual Information Flow 17

2.2 A Gaussian Lower Bound on the Mutual Information 19

2.2.1 Numerical Evaluation of the Gaussian Lower Bound 24

2.3 An Analytic Gaussian Lower Bound 26

2.3.1 Mean Field Approximation 26

2.3.1.1 Expectation of $\mathbf{\Lambda}_{h^{(\ell)}}$ 26

2.3.1.2 Expectation of $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$ 28

2.3.1.3 Lower Bound Approximation 29

2.4 Comparison of Mutual Information Approximation 31

2.5 Considerations on the Mean-Field Approximation 32

3 A Random Matrix Theory Approximation of the Mutual Information of Neural Networks at Initialisation 34
3.1 Introduction to Random Matrices . . . 35
3.1.1 Marchenko-Pastur Law . . . 36
3.1.1.1 Marchenko-Pastur Approximation for Mutual Information Computation . . . 37
3.1.2 The Stieltjes Transform . . . 40
3.1.2.1 Computing the Stieltjes Transform of the Product of Free Matrices . . . 42
3.1.2.2 Computing the Stieltjes Transform of the Sum of Free Matrices . . . 43
3.1.2.3 The Method of Moments . . . 44
3.1.3 Structure of Sections 3.2 and 3.3 . . . 46
3.2 Eigenspectrum of $\mathbf{\Lambda}_{h^{(\ell)}}$ . . . 47
3.2.1 Stieltjes Transform of the Matrix $\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(1)})\phi(\mathbf{h}^{(1)})^\top]$ with $\sigma_b = 0$ and $\sigma_n = 0$ . . . 48
3.2.2 Stieltjes Transform of the Matrix $\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})\phi(\mathbf{h}^{(\ell)})^\top]$ with $\sigma_b = 0$ and $\sigma_n = 0$ . . . 49
3.2.2.1 $2k$-cycle Contributions . . . 53
3.2.2.2 Agglomeration of the Contributions . . . 54
3.2.2.3 Proof of the Implicit Relation in Theorem 3.2.2 . . . 58
3.2.2.4 Numerical Results . . . 61
3.3 Eigenspectrum of $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$ . . . 61
3.3.1 Stieltjes Transform of the Matrix $\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(1)})\phi(\mathbf{h}^{(1)})^\top] - \mathbb{E}_x[\phi(\mathbf{h}^{(1)})\mathbf{x}^\top]\mathbb{E}_x[\mathbf{x}\phi(\mathbf{h}^{(1)})^\top]$ with $\sigma_b = 0$ and $\sigma_n = 0$ . . . 64
3.3.1.1 $2k$-cycle Contributions . . . 66
3.3.1.2 Agglomeration of the Contributions . . . 68
3.3.1.3 Proof of the Implicit Relation in Theorem 3.3.1 . . . 71
3.3.1.4 Numerical Experiments . . . 73
3.3.2 Stieltjes Transform of the Matrix $\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})\phi(\mathbf{h}^{(\ell)})^\top] - \mathbb{E}_x[\phi(\mathbf{h}^{(\ell)})\mathbf{x}^\top]\mathbb{E}_x[\mathbf{x}\phi(\mathbf{h}^{(\ell)})^\top]$ with $\sigma_b = 0$ and $\sigma_n = 0$ . . . 73
3.3.2.1 $2k$-cycle Contributions . . . 76

3.3.2.2 Aggregation of the Contributions and Implicit Relation in Theorem 3.3.5 . . . 81
3.3.3 Numerical Results . . . 82
3.4 Improvement on the Approximation of the Mutual Information Lower Bound . . . 85
3.4.1 Noise Application . . . 85
3.4.2 Estimation of the Mutual Information . . . 86
3.5 Considerations on the Eigen-distribution Approximation . . . 87

4 An Empirical Study of Derivative-Free-Optimisation Algorithms for Targeted Black-Box Attacks in Neural Networks 90
4.1 Adversarial Examples Formulated as an optimization Problem . . . 94
4.1.1 Derivative Free Optimization for Adversarial Examples . . . 96
4.1.2 Model-Based DFO . . . 97
4.1.2.1 BOBYQA . . . 98
4.2 Improving Efficiency and Computational Scalability . . . 98
4.2.1 Domain Sub-Sampling . . . 99
4.2.2 Hierarchical Lifting . . . 103
4.3 Comparison of Derivative Free Methods . . . 106
4.3.1 Parameter Setup for Algorithms . . . 107
4.3.2 Dataset and Neural Network Specifications . . . 109
4.3.3 Results for Standard and MadryLab Trained Neural Networks 112
4.3.4 Results with Fixed Pixel Count Constraints . . . 114
4.3.5 Relative Computational Cost . . . 115
4.4 Discussion and Conclusion . . . 117

5 Conclusion and Future Work 119
5.1 Information Propagation in a Random Neural Network . . . 119
5.2 Adversarial Examples . . . 120
5.3 Future Work . . . 121

A Proofs Relevant to the Method of Moments Applied to $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$ **122**

A.1 Proofs Relevant to Theorem 3.2.2 122

A.1.1 Initial Result 122

A.1.2 Proof of Lemma 3.2.3 - $E_2^{(n)}$ 128

A.1.3 Lemma 3.2.4 - $E_{2k}^{(n)}$ 130

A.1.3.1 Proof of Lemma 3.2.4 132

A.1.4 Proposition 3.2.5 135

A.1.4.1 Proof of Proposition 3.2.5 137

A.1.5 Side Theorems 144

A.2 Proofs Relevant to Theorem 3.3.1 148

A.2.1 Proof of Lemma 3.3.2 148

A.2.2 Lemma 3.3.3 149

A.2.2.1 Proof of Lemma 3.3.3 156

A.2.3 Proof of Proposition 3.3.4 157

A.2.3.1 Proof of Proposition 3.3.4 172

A.2.4 Side Theorems 173

A.2.5 Proof Corollary 3.3.1.1 175

A.3 Proofs Relevant to Theorem 3.3.5 175

A.3.1 Proof of Lemma 3.3.6 176

A.3.2 Lemma 3.3.7 177

A.3.2.1 Proof of Lemma 3.3.7 186

A.3.3 Lemma 3.3.8 187

B Experiments with Confidence Intervals **188**

Bibliography **191**

List of Figures

1.1 Flowchart of how the different AI systems relate to each other within AI itself according to [31]. Shaded boxes indicate components that are able to learn from data. 3
1.2 Example from [51] of how an image is processed through a neural network with the information flowing bottom up. At the bottom, the input image is decomposed in its different channels and then it is transformed with different features represented with the small rectangular images. Lower-level features act as oriented edge detectors while higher-level features are more abstract, and a score is computed for each image class in output. 5
1.3 Example of how an adversarial attack on an image of a squirrel (a) classified correctly, to which an imperceptible perturbation (b) is added, is then classified (c) as a parking meter. 6

2.1 Illustration of the ordered-chaos transition when $\phi(h) = tanh(h)$ verifying the work in [88]. The critical line $\chi = 1$ defines the boundary between the two phases. In the chaotic regime $\chi > 1$ and gradients explode in expectation, while in the ordered regime $\chi < 1$ and gradients vanish in expectation. 16
2.2 Figure reproduced from [91] illustrating how the mutual information between the input and the hidden layers, and between the hidden layers and the output evolve during training for different training dataset dimension (5% of the data on the left, 45% in the middle, and 85% on the right). The number of training epochs for each state is described by its colour. For more details see [91] 17

2.3 Figure reproduced from [1] depicting the convergence of the mutual information between the input $\mathbf{x}$ and the post-activation hidden layer $\mathbf{h}^{(\ell)}$ for a random neural network with $\sigma_b = 0$ and $\phi(\cdot) = Htanh(\cdot)$[1]when σ_w varies. 18
2.4 Signal Propagation in the perturbed neural network introduced in [30]. 20
2.5 Numerical approximation by sampling of the Gaussian lower bound in (2.48) for the conditional mutual information of a feed-forward neural network when σ_w varies along the edge of chaos with $n = 90$ and $\phi(\cdot) = tanh(\cdot)$. 25
2.6 In (a) and (b) we compare mean field approximation of the mean value on the diagonal of the matrices $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^{\top}\mathbf{\Sigma}_{xh^{(\ell)}}$ respectively with the empirical distribution in function of the layers. The empirical distribution is obtained by considering 100 different set of weights $\mathcal{W}^{:\ell}$ for the neural network when $n = 50$ and with 10^5 different random inputs. The shading of the distribution corresponds to the 50%, 80%, and 100% confidence intervals in transparency order. 27
2.7 Comparison of the mean field approximation of the Gaussian lower bound $I^{\mathcal{G}}(\mathbf{x}, \mathbf{h}^{(\ell)})$ in (2.47) and its sampled value for a neural network with $n = 90$ and $tanh(\cdot)$ activation function when the variance σ_w is changed. Here σ_b is chosen so that (σ_w, σ_b) satisfies the EoC condition. 29
2.8 Comparison between the empirical lower bound and the one obtained with mean field approximation for a neural network with $n = 1000$ when the σ_w is changed on the x axis and σ_b is kept fixed to 0 with $\sigma_n = 0.1$. 30
2.9 Comparison of the sampled Gaussian lower bound $I^{\mathcal{G}}(\mathbf{x}, \mathbf{h}^{(\ell)})$ and the analytic approximation in (2.59) for $n = 30, 60$, and 90 for layers 2 (a) and 18 (b). 30

2.10 Comparison between the approximation of $I(\mathbf{x}, \mathbf{h}^{(\ell)})/n$ with the replica method (Replica Mthd) in [28] and the lower bound obtained with mean field approximation (MF Apprx) for a neural network with $n = 1000$ when the σ_w is changed on the x axis and σ_b is kept fixed. The results relative to this graph consider the post-activation hidden layer rather than the pre-activation hidden layer. 32
2.11 Convergence of the lower bound $I^{\mathcal{G}}(\mathbf{x}, \mathbf{h}^{(\ell)})/n$ obtained with the mean field approximation when we change the variance σ_w and keep $\sigma_b = 0$ and $\sigma_n = 1e-4$. 33
2.12 Lower bound $I^{\mathcal{G}}(\mathbf{x}, \mathbf{h}^{(28)})/n$ obtained with the mean field approximation when changing the variances (σ_w, σ_b). The dashed lines correspond to the maximum mutual information lower bound once σ_b is considered fixed. 33

3.1 Example of how the Marchenko-Pastur distribution in Theorem 3.1.1 describes the eigen-distribution of a sample covariance in practice when $\gamma = 0.5$, $\sigma^2 = 1$, and $n = 100$. 37
3.2 Examples of the distribution of the eigenvalues of $\mathbf{\Lambda}_{h^{(\ell)}}$ (left in blue) and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^{\top}\mathbf{\Sigma}_{xh^{(\ell)}}$ (right in green) at the 3rd (Above) and 7th layer (below). The dash red line corresponds to the Marchenko-Pastur distribution with the same mean as the spectrum. 38
3.3 Admissible patterns according to Definition 3.1.1 when considering k=3. Any admissible graph with $k = 3$ consist of a permutation of the indices in the illustrated patterns. This Figures reproduces parts of [72, Figure S1]. 46
3.4 Approximation of the empirical eigen-spectrum for the matrix $\mathbf{\Lambda}_{h^{(\ell)}}$ with $n = 100$ and $(\sigma_w, \sigma_b, \sigma_n) = (2, 0, 0.001)$ at the first four layers. The Stieltjes transform induced approximation models the tail of the distribution more accurately than the Marchenko-Pastur approximation does, especially for the second layer where the approximation seems to fit the empirical distribution exactly. 50

3.5 Comparison of the Wasserstein distance between the empirical distribution of $\mathbf{\Lambda}_{h^{(\ell)}}$ for different network dimensions and the distribution attained with the Stieltjes transform (dash-dotted) and with the Marchenko-Pastur distribution (solid). For the second layer the Wasserstein distance decreases when n increases as expected. . . . 51
3.6 Example on how the contribution of two graphs G_1 and G_2 determines the contribution of their combination G_3 with a μ identification. The expressions of the marginal contributions is related to Proposition A.1.4. Note, with an I_i identification, we would have obtained that $\mathfrak{m}_5^{G_3} = \mathfrak{m}_3^{G_1} n_1/n_1 \mathfrak{m}_2^{G_2}$. 57
3.7 Illustration of the contributions that have to be considered for different cycles of dimension ℓ. This Figure reproduces [6, Figure 14]. . 61
3.8 Approximation of the eigenspectrum for the matrix $\mathbf{\Lambda}_{h^{(\ell)}}$ with $n = 100$ and $(\sigma_w, \sigma_b, \sigma_n) = (2, 0, 0.001)$ for six different layers. The histogram (Empirical Results) corresponds to the eigenvalues obtained on 100 different simulations and it is compared to the Marchenko-Pastur distribution with the same mean, to the distributions obtained with the Stieltjes transform of Theorem 3.2.1 (Stilejes Transform) and of Theorem 3.2.2 (Stieltjes Transform Covariance). 62
3.9 Comparison of the Wasserstein distance between the empirical distribution of $\mathbf{\Lambda}_{h^{(\ell)}}$ for different network dimensions and the distribution attained with the Stieltjes transform (dash-dotted), with the Stieltjes transform with correlated inputs (dashed) and with the Marchenko-Pastur distribution (solid). Among the approximations, the one considering the correlation of the inputs is the only to not be affected by the depth of the layer suggesting that it is able to approximate the distribution appropriately at any layer. 63
3.10 Empirical distribution of the eigen-spectrum for the matrices $\mathbf{\Lambda}_{h^{(2)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(2)}}^\top \mathbf{\Sigma}_{xh^{(2)}}$ and $\frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(2)}}^\top \mathbf{\Sigma}_{xh^{(2)}}$ with $n = 100$ and $(\sigma_w, \sigma_b, \sigma_n) = (2, 0, 0.001)$. 74
3.11 Empirical distribution of the eigenspectrum for the matrix $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$ with $n = 100$ and $(\sigma_w, \sigma_b, \sigma_n) = (2, 0, 0.001)$ at the first four layers. 83

3.12 Comparison of the Wasserstein distance between the empirical distribution of $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^{\top}\mathbf{\Sigma}_{xh^{(\ell)}}$ for different network dimensions and the distribution attained with the Stieltjes transform (dash-dotted) and with the Marchenko-Pastur distribution (solid). For the second layer the Wasserstein distance decreases when n increases as expected. The modelling of deeper layers has a systematic error that is not accounted for and therefore the error is not decreasing with n. Further, the Marchenko-Pastur distribution does not converge to the empirical distribution with n for any layer. 84

3.13 Comparison of the empirical mutual information lower bound between the input and the second hidden layer with its approximation with the mean-field approximation and with the Stieltjes transform. In (a) the results are relative to the noise $\sigma_n = 0.1$, while in (b) to the noise $\sigma_n = 0.01$. In (b) the empirical mutual information could not be computed for small σ_w since the matrices were too badly conditioned. 87

3.14 Comparison of the empirical mutual information lower bound between the input and the third hidden layer with its approximation with the mean-field approximation and with the Stieltjes transform. In (a) the results are relative to the noise $\sigma_n = 0.1$, while in (b) to the noise $\sigma_n = 0.01$. In (b) the empirical mutual information could not be computed for small σ_w since the matrices were too badly conditioned. 88

4.1 The success rate of targeted attacks as a function of the perturbation's allowed ℓ^∞ magnitude for algorithms: GenAttack [4], Parsimonious [60], Square [5], Frank-Wolfe [12], and the BOBYQA based algorithm introduced here. Specifically for a ResNet50 network trained either on (a) the CIFAR10 or (b) the ImageNet dataset with (Adv) and without (Non-Adv) the defense by MadryLab [24]. An attack is considered successful if the method found the targeted adversarial example with less than 3,000 or 15,000 queries to the network trained on CIFAR and ImageNet dataset, respectively; Results for the case where the success rate is null, i.e. when no perturbations were successful, are excluded from the plot. 92

4.2 Example of how the perturbation $\boldsymbol{\eta}$ evolves through the iterations when an image in $\mathbb{R}^{4\times4}$ is attacked. In (a) the perturbation is $\boldsymbol{\eta} = \boldsymbol{\eta}^0$ and a sub-domain of $b = 4$ pixels (in red) is selected. Once the optimal perturbation $\boldsymbol{\eta}^1$ in the selected sub-domain is found, the perturbation is updated in (b) and a new sub-domain of dimension b is selected. The same is repeated in (c). 100

4.3 Cumulative distribution function of successfully perturbed images as a function of number of queries by the BOBYQA based algorithm attacking neural networks trained on the MNIST and the CIFAR10 datasets. In each image the effectiveness of different sub-sampling methods in generating a successful adversarial example is shown for different values of maximum perturbation energies ε_∞. See [103] for details about experimental setup. 101

4.4 Impact of hierarchical lifting approach on Loss function (4.2) as a function of the number of queries to a ResNet50 trained on ImageNet dataset to find the adversarial example for a single image with the BOBYQA based method. The green vertical lines correspond to changes of hierarchical level, which entail an increase in the dimension of the optimization space. 102

4.5 Example of how the perturbation $\boldsymbol{\eta}$ is generated in a hierarchical lifting method with $n_1 = 4$ and $n_2 = 16$ on an image in $\mathbb{R}^{12\times 12}$. In (a) the perturbation is $\boldsymbol{\eta} = \boldsymbol{\eta}_0$ and the boxes generated via the grid of dimension n_1 are highlighted in red. Once the optimal perturbation $\boldsymbol{\eta}_1$ is found, the perturbation is updated in (b) and the image is further divided with a grid with n_2 blocks. The final solution obtained after optimization is shown in (c). 104

4.6 Examples for (a) random and (b) block liftings. In the random case each pixel in the perturbation is associated to just one element of $\hat{\boldsymbol{\eta}}_\ell$. Block lifting uses a piece-wise constant interpolation $\mathbf{L}$ over a coarse grid $\mathbf{S}\hat{\boldsymbol{\eta}}_\ell$ and each block is associated uniquely to one of the variables in $\hat{\boldsymbol{\eta}}_\ell$. In both cases, the lifting $\mathbf{D}$ is such that each element $\mathbf{D}_{ij}$ is either 1 or 0. 105

4.7 Cumulative fraction of test set images successfully misclassified with adversarial examples generated by GenAttack, Parsimonious, Square, Frank-Wolfe, and our BOBYQA based approaches for different maximum perturbation energies ε_∞ and neural networks trained on the CIFAR10 dataset. In all results the solid and dashed lines denoted by 'Non-Adv' and 'Adv' corresponds to attacks on networks trained without or with the MadryLab defense strategy [24] respectively. . 110

4.8 Cumulative fraction of test set images successfully misclassified with adversarial examples generated by GenAttack, Parsimonious, Square, Frank-Wolfe, and our BOBYQA based approaches for different maximum perturbation energies ε_∞ and neural networks trained on the ImageNet dataset. In all results the solid and dashed lines denoted by 'Non-Adv' and 'Adv' corresponds to attacks on neural networks trained without or with the MadryLab defense strategy [24] respectively. 111

4.9 Cumulative fraction of test set images successfully misclassified with adversarial examples generated by Parsimonious, Square, and our BOBYQA based approaches for different maximum perturbation energies ε_∞ against a ResNet50 trained non-adversarially on the CIFAR10 dataset when only the 1000 pixels with the highest variance in intensity in their neighborhood are allowed to be modified. 116

A.1 Example of two non-admissible graphs with respectively $k = 6$ on the left, and $k = 4$ on the right. 148

A.2 Example of how in the Schur complement the contribution to the momentum is defined on as many independent blocks of products, Π_j, as the number of factors in the form $\sigma_x^2 \mathbf{W}_{i_\xi p} \mathbf{W}_{i_{\xi+1} p} \theta_3^2$. 149

A.3 Diagram of sequences of paths that allow to connect i_1 to i_{l+1}, with $l = 5$, for different number of steps k. The blue line corresponds to the unique path connecting i_0 to i_l with $k = l$ steps; the dashed magenta lines corresponds to the l paths connecting i_0 to i_l with $k = l + 2$ steps; and the dotted red lines corresponds to the l^2 paths connecting i_0 to i_l with $k = l + 4$ steps. 153

A.4 Illustration of some examples of how the $\sigma_x^2 \theta_3^2 \mathbf{W}_{i_\xi p} \mathbf{W}_{i_{\xi+1} q}$ factors might be arranged around some identifications. 160

B.1 Cumulative fraction of test set images successfully misclassified with adversarial examples generated by GenAttack, Parsimonious, Square, Frank-Wolfe, and our BOBYQA based approaches for different maximum perturbation energies ε_∞ and DNNs trained on the CIFAR10 dataset. In all results the solid and dashed lines denoted by 'Non-Adv' and 'Adv' corresponds to attacks on networks trained without or with the MadryLab defense strategy [24] respectively. The shaded region corresponds to the 90% confidence intervals for each method computed via boot-strapping the data. 189

B.2 Cumulative fraction of test set images successfully misclassified with adversarial examples generated by GenAttack, Parsimonious, Square, Frank-Wolfe, and our BOBYQA based approaches for different maximum perturbation energies ε_∞ and DNNs trained on the ImageNet dataset. In all results the solid and dashed lines denoted by 'Non-Adv' and 'Adv' corresponds to attacks on networks trained without or with the MadryLab defense strategy [24] respectively. The shaded region corresponds to the 90% confidence intervals for each method computed via boot-strapping the data. 190

Chapter 1

Introduction

Over the last decade, the academic and industrial communities have become increasingly involved in the field of **Artificial Intelligence** (AI), especially focusing on **Machine Learning** (ML); systems that are able to gain generalising knowledge about a task by extracting patterns from raw data [31]. The initial interest in the field was sparked by the human-like performances of ML algorithms in different challenges such as image recognition [84]. These tasks are easy for humans to execute but difficult to describe formally, implying that successful ML systems can infer relevant non-trivial features of a problem; this suggested that more complex problems than image recognition could have been tackled. The resulting self-perpetuating dynamic, where practical and theoretical results would allow one-another to develop, lead during the last years to surprising results. ML based algorithms achieved state-of-the-art performances in strategy games [92, 105], creative content generation [15, 108, 109], translation [78], and medical applications such as protein folding [46] and causal inference on carcino-genesis [67].

The effectiveness of ML algorithms strongly depends on the structure of the data. Many tasks can be solved via machine learning algorithms if they are provided with the right features describing the input data. Figure 1.1 shows a summary of how the different AI systems relate to each other according to [31]. In **classical ML** the features are hand-designed from the inputs and the algorithms learn a mapping directly from the features. However, it is difficult to know *a priori* what set of features should be considered. For example, to detect a dog in an image one good candidate for a defining feature could be the tail or the snout. However,

extracting such a feature might not be trivial because of how the animal is facing the camera, or how intense the daylight is (among many other variables). **Representation learning** uses the data itself to directly learn the relevant features for the given problem [31]. This approach has the advantage of not being dependent on the task itself and of being time efficient compared to manual extraction of the features. Moreover, these algorithms can extract and use features that are not easily hand-constructed, allowing a further level of abstraction.

Deep Learning (DL), one type of representation learning, is particularly good for learning different levels of feature abstraction and how the features relate to one another. It looks for a hierarchical relation among features, starting with learning simple, elementary features, and learning how they combine and relate to one another to construct more complex features. This structure was inspired by how brains are thought to learn; "The standard model of the visual cortex suggests that the brain first extracts edges, then patches, then surfaces, the objects, etc." [62]. As a consequence, DL algorithms are also known as **Artificial Neural Networks**, or simply **Neural Networks**.

1.1 Neural Networks

A Neural Net F is a function which maps an input space $\mathcal{X}$ to an output space $\mathcal{Y}$, $F : \mathcal{X} \to \mathcal{Y}$, and is parameterised by a set of parameters θ. The values of these parameters θ are adjusted during a training stage, which attempts to minimise a loss function measuring the accuracy of the net's prediction over a given set of data. Feed-forward Neural Networks are the building blocks of DL and consist of an iterative linear operation over the input to which a nonlinear function is applied element-wise. The final output is given as an input to a simple operator specific to the application of the net; for example, in image classification the final output is given to an `argmax` function which identifies the class of an image.

In a feed-forward neural network with L layers of dimension n^k with $k \in \{1, \ldots, L\}$, the input $\mathbf{x}^{(0)} \in \mathcal{X} \subset \mathbb{R}^{n_0}$ is processed by the following iterative operation

$$\mathbf{x}^{(k)} = \phi^{(k)} \left(\mathbf{W}^{(k)} \mathbf{x}^{(k-1)} + \mathbf{b}^{(k)} \right) \quad \text{for } k = 1, ..., L \tag{1.1}$$

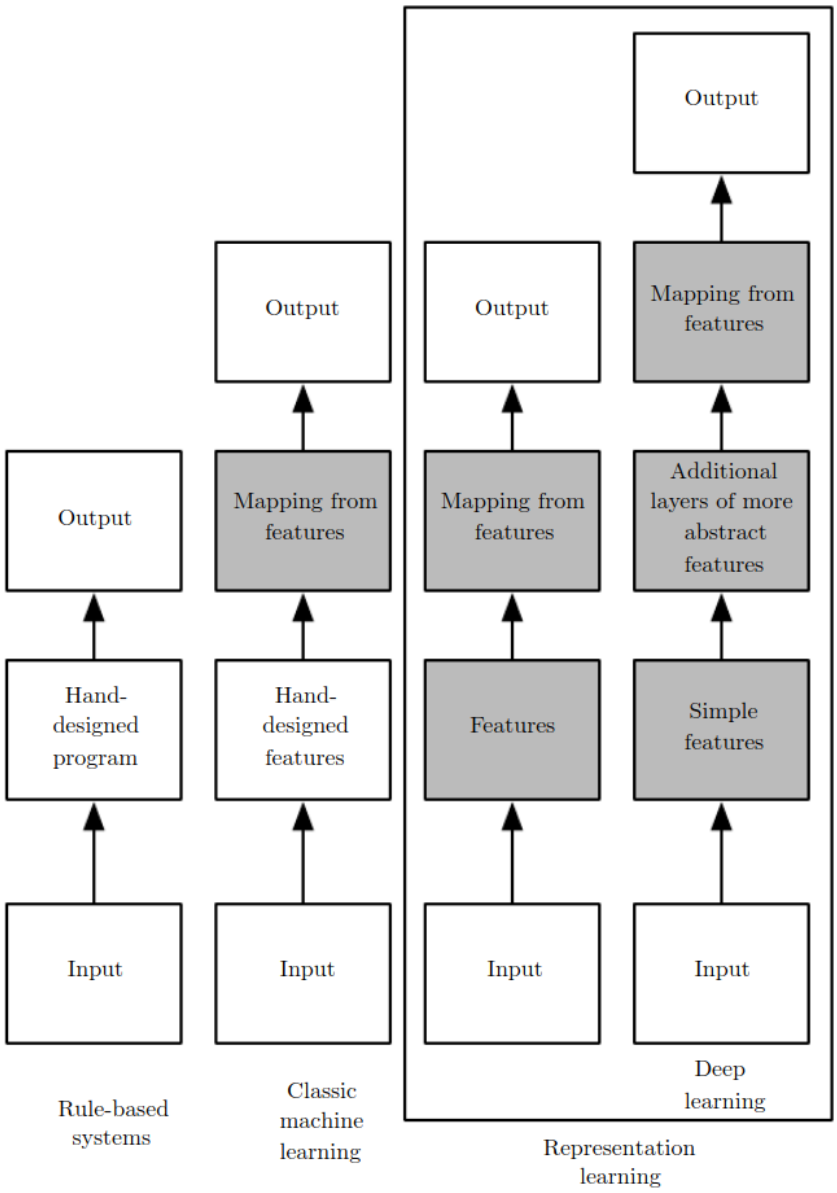


Figure 1.1: Flowchart of how the different AI systems relate to each other within AI itself according to [31]. Shaded boxes indicate components that are able to learn from data.

where $\phi^{(k)}(\cdot)$ is an element-wise nonlinear activation function such as $tanh(\cdot)$[1], and $\mathbf{W}^{(k)} \in \mathbb{R}^{n_{k+1} \times n_k}$ and $\mathbf{b}^{(k)} \in \mathbb{R}^{n_{k+1}}$ are the weights of the iteration, or layer, k. The parameters of the feed-forward neural network are the set of weights, i.e. $\theta = \{\mathbf{W}^{(k)}, \mathbf{b}^{(k)}\}_{k=1}^{L}$. The variable $\mathbf{x}^{(k)}$ is sometimes also referred to as hidden layer. In the following sections we will also consider the pre-activation hidden layer $\mathbf{h}^{(k)} = \mathbf{W}^{(k)}\mathbf{x}^{(k)} + \mathbf{b}^{(k)}$, also written in function of its input, e.g. $\mathbf{h}^{(k)}(\mathbf{x}^{(0)})$, and we are going to refer to feed-forward neural networks by neural networks.

Within the topics concerning neural networks, this thesis is relevant to their initialisation and to their robustness.

1.1.1 Initialisation of Neural Networks

This apparently simple structure has potentially unlimited approximation power. The **universal approximation theorem** developed by Hornik et al. [39] and Cybenko [18] showed that a one-layer neural network can approximate any Borel measurable function provided that the neural network has enough hidden units. Subsequent research also showed that adding depth to a neural network allows better expressivity than adding the same number of parameters in width [100, 110]. This was observed also in practice [93], along with the fact that features tend to be more abstract the deeper the layer in the neural network, see Figure 1.2.

Although theoretically there is a set of parameters such that a neural network is able to approximate a given function, in practice it is necessary to initialise the weights of the neural network with an initial guess which might yield inaccurate predictions. To obviate this issue, they are usually chosen at random and then modified to find this optimal neural network using a loss function. This procedure is referred to as the training of the neural network. In the training stage we are given a set of m labelled data $\{(\mathbf{x}_{tr}^i, \mathbf{y}_{tr}^i)\}_{i=1}^m$, with $\mathbf{x}_{tr}^j \in \mathcal{X}$ and $\mathbf{y}_{tr}^j \in \mathcal{Y}$ for each j. Define $\hat{\mathbf{y}} := F(\mathbf{x}; \theta)$ to be the approximation of the neural network to the label $\mathbf{y}$ of the input $\mathbf{x}$, and an error measure, $L(\mathbf{x}, \mathbf{y}, \hat{\mathbf{y}})$, of how accurate the approximation $\hat{\mathbf{y}}$ is. Since $\hat{\mathbf{y}}$ is a function of the input and the parameters of the neural network, i.e. $\hat{\mathbf{y}} =: \hat{\mathbf{y}}(\mathbf{x}, \theta)$, the error function is noted as $L(\mathbf{x}, \mathbf{y}, \theta)$. For example, if the output is a vector, the mean square error could be considered as an error measure,

[1] In this thesis we are going to consider the same activation function ϕ for all the layers.

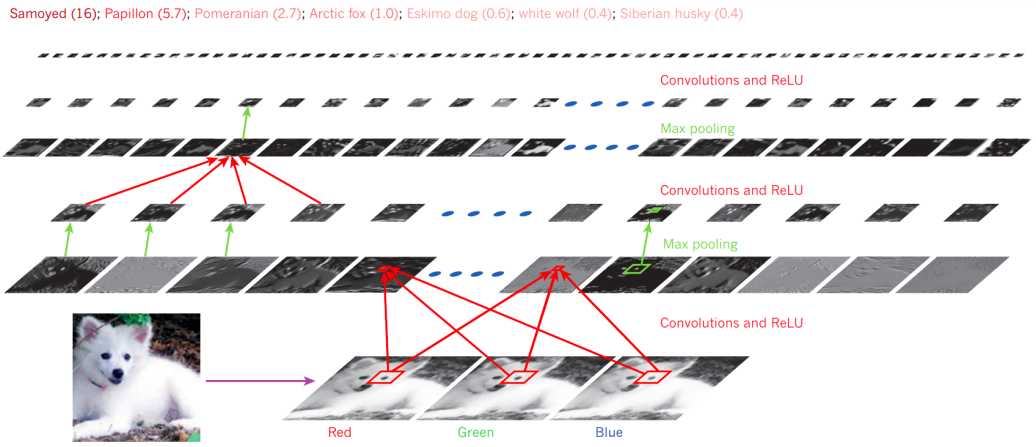


Figure 1.2: Example from [51] of how an image is processed through a neural network with the information flowing bottom up. At the bottom, the input image is decomposed in its different channels and then it is transformed with different features represented with the small rectangular images. Lower-level features act as oriented edge detectors while higher-level features are more abstract, and a score is computed for each image class in output.

$L(\mathbf{x}, \mathbf{y}, \theta) = \sum_i (\hat{\mathbf{y}}(\mathbf{x}, \theta)_i - \mathbf{y}_i)^2$. The parameters θ are obtained by minimising the total error measure $\mathfrak{L}(\theta) = \sum_{i=1}^{m} L(\mathbf{x}_{tr}^i, \mathbf{y}^i, \theta)$ over the whole dataset in the problem,

$$\theta = \arg\min_{\eta} \mathfrak{L}(\eta). \tag{1.2}$$

The optimisation of the parameters usually follows traditional optimisation methods that are based on local information such as the gradient or the Hessian of the loss function. While the computation of the gradient can be efficiently computed with the backprop algorithm [83] allowing efficient implementations of gradient based methods [8], the computation of the Hessian is too computationally demanding and therefore is usually approximated [7, 56]. Given that the loss function is generally non-convex, these local methods might get stuck in local minima [65].

In this context, the problem of how to initialise the neural networks becomes relevant since the way the parameters are randomly sampled affects the behaviour of the optimisation procedures. For example, the convergence to a good solution might be thwarted by numerical inaccuracies arising because of the magnitude of the gradients begin too big or too small. To avoid these kind of problems, different



(a) Original image. (b) Perturbation. (c) Modified image

Figure 1.3: Example of how an adversarial attack on an image of a squirrel (a) classified correctly, to which an imperceptible perturbation (b) is added, is then classified (c) as a parking meter.

approaches have been identified suggesting which initialisation would be the easiest to train from [74, 88].

1.1.2 Adversarial Examples

It has been shown that neural networks are susceptible to adversarial examples [32, 98]. In image recognition tasks, we may find tiny perturbations which, when added to an image, lead a neural network to misclassify it although correctly classifying it while unperturbed. An example is shown in Figure 1.3.

This susceptibility has profound implications for these algorithms in application to real world problems. For example, autonomous vehicles could be led to ignore stop signs and pedestrians [19, 25, 49, 94, 111]. Without having access to how the lane-control algorithm of Tesla's self driving system works, Tencent Keen Security Lab was able to have the car perceiving a non-existing lane by perturbing few pixels of the front camera potentially causing an unwanted lane switch[2].

Since the discovery of these vulnerabilities, there has been a lot of research focusing on finding the most efficient adversarial attacks in different settings, such as the black-box one considered in the example above where the attacker cannot access any inofrmation of the model attacked. There are multiple purposes for this,

[2]The report on this attack is available at `https://keenlab.tencent.com/en/whitepapers/Experimental_Security_Research_of_Tesla_Autopilot.pdf`

as reported in [89, 111]; it contributes to understanding more deeply what kind of neural networks we are working on, generating a measure of the robustness of these algorithms, and finding successful defensive strategies with which to generate future neural networks. See [81] for examples on how defences are being developed keeping into consideration the effectiveness of the different attacks.

Following the notation in [113], these perturbed images that are misclassified by the neural networks, are defined as *adversarial examples*, and they lead to the definition of two different accuracy measures for a trained neural network. The *natural* accuracy, or *natural risk*, refers to when non-adversarial inputs are given to the neural network; *robust accuracy*, on the other hand, considers only adversarial examples as an input. These adversarial perturbations arise from the fact that neural networks are essentially partitioning the domain, and the decision boundaries are defined by maximising the natural accuracy uniquely. This implies that the neural networks are not trained to also correctly approximate the output for the neighbourhood of the data [112].

1.2 Research Objectives and Contributions

This thesis aims at exploring two general questions relative to neural networks.

- How does information theory relate the initialisation of the weights within the neural network to the propagation of signals within the neural network itself?

- How robust are neural networks to adversarial misclassification in the setting where the network is only accessible through the output of as few network input output pairs as possible?

Regarding the initialisation of neural networks, we considered an information theoretic perspective following some initial results in [1]. Information theory studies how signals propagate in noisy channels and we relied on its mutual information measure to study how at initialisation the choice of the weights' distribution affects the information that each hidden layers has of the input. In [99] we introduced a lower bound for the mutual information of a random neural network and a coarse

analytical approximation. In our latest work included in Chapter 3, we focused on improving the approximation to the lower bound of the mutual information by considering tools form random matrix theory.

In the robustness setting, we limited our research to black-box adversarial examples where an attacker has no information on the neural network beside its output. In [103] we introduced a novel adversarial-example algorithm based on the BOBYQA method [80], a common technique in the optimisation literature. Later on, in [104] we explored for the first time how the different methods compare in practice on the same exact settings and found out that our method achieves state-of-the-art performance in certain conditions. Specifically we highlighted the intuitive property that adversarial methods are not universally efficient; the more complex a setting is, the more sophisticated an adversarial example method should be.

1.3 Thesis Structure

Chapters 2 and 3 are going to delve into the question of how information theory can enhance our understanding of neural networks at initialisation.

Specifically, in Section 2.1 we start by introducing random neural networks, which model neural networks at initialisation, and by summarising the research that was conducted to understand the geometric flow, the gradient flow, and the mutual information flow within them. Specific to this last topic, the work in [1] empirically suggests that some initialisation preserve information propagation through the layers. However, these analyses are based on mutual information estimation methods that are not always consistent and do not give an insight into what preserves information propagation. This contextualisation allows to then introduce in the remainder of the chapter the work presented at the 2021 IEEE International Seminar on Information Theory by J. Tanner and G. Ughi [99]. In Section 2.2 we introduce a lower bound of the mutual information, that is numerically computable, and we show in Section 2.3 that we can approximate it with an analytical formula that is based on the mean-filed approximation. As we then show in Section 2.4 this analytical approximation is qualitatively accurate and supports the same ideas of some initialisations that propagate more information

through the layers. Furthermore, since the computation of the analytical lower bound is time efficient, it is possible to run experiments for very deep networks and the results suggest that optimal initialisations from a mutual information perspective do sometimes coincide with those that are optimal from a gradient or geometric perspective.

In Chapter 3 we improve the approximation of the lower bound introduced in Section 2.2 by relying on tools from random matrix theory. We start in Section 3.1.1 by giving an introduction of some standard results in random matrix theory, namely the Marchenko-Pastur distribution, which yield an approximation of the lower bound in Section 2.2 that coincides with the analytical expression in Section 2.3. This leads to the introduction in Section 3.1.2 of the Stieltjes transform, a sophisticated technique that allows for improvement in the estimation of the lower bound thanks to its properties.

The improvement in the approximation is a consequence of how the Stieltjes transform allows to approximate more precisely the eigen-distribution of the two matrices $\Lambda_{h^{(\ell)}}$ and $\Lambda_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\Sigma_{xh^{(\ell)}}^\top \Sigma_{xh^{(\ell)}}$ whose log-determinant defines the lower bound in Section 2.2. In Section 3.2 we first present the Theorem from [72] that allows computing the exact asymptotic eigen-distribution of $\Lambda_{h^{(\ell)}}$ at the first layer in Section 3.2.1. We expand on this work in Section 3.2.2 proving a theorem defining the exact asymptotic eigen-distribution of $\Lambda_{h^{(\ell)}}$ for deeper layers. In Section 3.3 we follow a similar structure as in Section 3.2; we first prove in Section 3.3.1 a Theorem that allows computing the exact asymptotic eigen-distribution of $\Lambda_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\Sigma_{xh^{(\ell)}}^\top \Sigma_{xh^{(\ell)}}$ at the first layer, and we expand this to deeper layers in Section 3.3.2 though without being able to retrieve an exact asymptotic expression of the eigen-distribution. In Section 3.4 the Theorems in Sections 3.2 and 3.3 are used to approximate the lower bound and an improvement in the accuracy of the approximation is observed compared to the mean-field based one. However, as stated in the conclusive remarks in Section 3.5, because of how time demanding it is to compute these approximations, it is possible to consider only shallow networks. Nevertheless, thanks to the modelling of the eigen-distribution in Sections 3.2 and 3.3 we are able to identify a variable that drives the reduction in mutual information; thus gaining an understanding that the previous work was lacking.

In Chapter 4 the focus is shifted towards the study of the robustness of neural networks against adversarial misclassifications and is based on the work published in the "Optimization and Engineering" Springer Journal by G. Ughi, V. Abrol, and J. Tanner [104]. After a brief description of how adversarial methods are classified, in Section 4.1 the generation of adversarial examples is associated with the solution of a constrained optimisation problem. Since the work focuses on black-box adversarial examples, the optimisation algorithms used to generate adversarial examples cannot rely on gradient samples. In Section 4.1.1 we review the derivative-free-optimisation methods that have been used in previous literature to generate adversarial examples. In Section 4.1.2 the noticeably missing model-based derivative-free-optimisation methods are introduced and more details are given about the BOBYQA method since we build on it to generate an algorithm with the aim to have a complete perspective of how different derivative-free-algorithms perform in generating adversarial examples. Any derivative-free-optimisation method struggles when the dimensionality of the problem increase, and because of the high-dimensional nature of inputs to neural networks the algorithms used for adversarial examples generation all rely on some dimensionality reduction techniques; these are all summarised in Section 4.2 in a unifying formulation. In Section 4.3 all the different introduced algorithms to generate adversarial examples are compared for the first time in exactly the same two different experimental frameworks. As is further discussed in Section 4.4, the experimental results suggest that combinatorial algorithms work generally better and that there is a lot of variation in the performances when the settings change, therefore any new adversarial algorithm should be compared in many settings.

In the conclusive Chapter 5, a summary of the results achieved in this thesis is going to be presented, and future research directions are going to discussed.

1.4 Collaboration with New Rock Management

The research conducted in this DPhil is partially funded by New Rock Capital Management (NRCM), now RE5Q, through the InFoMM program. NRCM is a firm focused on real estate analytics, portfolio analysis, and risk management.

They are interested in using publicly available data and machine learning techniques to develop a computerised appraisal system, from which fast, objective, and accurate evaluations can be obtained. Such a system would revolutionise the current market based on commercial real estate evaluation, since this task is currently reliant upon a subjective review undertaken by an agent which is slow, labour intensive, and expensive.

The implementation of such a system requires also the development of a method able to determine the quality of the interiors of the considered properties. To do this, we applied deep learning algorithms to the available images of the estate obtaining some interesting results. The main concern with this approach however was that the appraisal of an estate through an NRCM platform might be tampered with by perturbing the images. Therefore, the need of understanding how susceptible the appraisal platform could have been, led to the work on adversarial examples presented in Chapter 4.

The experiments showed that adversarial attacks should not be a threat to NRCM and this allowed us to focus on how to improve the training of neural networks that can be used for appraisal algorithms. Abrol and Tanner [1] suggested that it is possible to improve the training by finding an optimal initialisation from a mutual information point of view. Therefore, we focused on measuring this metric at initialisation. Unfortunately, we did not have enough time to validate the improvement in performance of neural networks initialised optimally from a mutual information point of view.

Chapter 2

A Mean-Field Approximation of the Mutual Information of Neural Networks at Initialisation[1]

Random neural networks are random nonlinear functions from which neural networks are sampled, determining their initialisation, and subsequently trained to map a training set of inputs to known outputs. A given distribution of random neural networks, i.e. a given set of initialisations, is determined by $(\sigma_w, \sigma_b, \phi(\cdot))$ where $\phi(\cdot)$ is the activation function and σ_w and σ_b define the Gaussian distributions from which the weights are sampled according to

$$\mathbf{W}_{ij}^{(\ell)} \sim \mathcal{N}(0, \sigma_w^2/n_\ell) \quad \text{and} \quad \mathbf{b}_i^{(\ell)} \sim \mathcal{N}(0, \sigma_b^2), \tag{2.1}$$

with n_ℓ being the width of layer ℓ. This kind of setting for neural networks initialisation was first introduced in [29].

Different work has been conducted to understand how the choice of $(\sigma_w, \sigma_b, \phi(\cdot))$ affects the sampled neural networks. Most notably, the work in [88] first showed that neural network initialisations that preserve information about the inputs are typically easier to train. This conclusion was based on geometric considerations of input signal dynamics in a random neural network, measured with the distributions of intermediate hidden layers. These kind of analysis were able to suggest for a given nonlinear activation $\phi(\cdot)$ an optimal set of initialisation parameters (σ_w, σ_b).

[1]The material in the Chapter is primarily taken from the work presented at the 2021 IEEE International Seminar on Information Theory [99]

Here we conduct an alternative information theoretic investigation of random feed-forward neural networks in order to determine how the choice of $(\sigma_w, \sigma_b, \phi(\cdot))$ impacts the flow of information through the neural network. We do this by considering the mutual information relative to the inputs and their associated hidden layers $(\mathbf{x}, \mathbf{h}^{(\ell)}) \in \mathcal{X} \times \mathcal{H}^{(\ell)}$, a measure of the dependence of two random variables. By deriving a lower bound of the mutual information as a function of $(\sigma_w, \sigma_b, \phi(\cdot))$, we suggest that initialisations which are optimal for geometric training analysis are also preferable from a mutual information perspective.

In Section 2.1, we give an introduction to the results relevant to random neural networks. This also allows to introduce some results relevant to the mutual information in neural networks. In Section 2.2, we prove the existence of a lower bound on the mutual information problem by introducing the Gaussian model of the neural network. In Section 2.3, we approximate this lower bound via the mean field approach. Finally in Section 2.4, we compare our lower bound approximation to a state-of-the-art mutual information approximator and see that the initialisation at the edge of chaos similarly increases the mutual information at deep layers.

2.1 Introduction to Random Neural Networks

The study of random neural networks became central in the literature on how to initialise neural networks, as it allowed the understanding of how the weights' distribution affects the expressivity and the trainability of a neural network.

Therefore, we first present the main results relevant to the geometric flow and to the gradient flow which respectively give some suggestions on how expressive and how trainable a neural network is at initialisation. Hereafter, we will consider how mutual information flow can be inserted in this initialisation context.

2.1.1 Geometric Flow

Poole et al. [76] introduced the mean-field approach which consists of assuming that with randomly distributed weights each pre-activation layer is distributed according to a Gaussian in the large n limit. This approximation allows one to study the expected norm of a hidden layer and the expected covariance of two hidden layers generated by two different inputs.

The norm of the input at layer ℓ, is defined as

$$q^{(\ell)} = \frac{1}{n_\ell} \sum_j \left(\mathbf{h}_j^{(\ell)} \right)^2, \tag{2.2}$$

which corresponds to the second moment of the pre-activation layer $\mathbf{h}^{(\ell)}$. Assuming that in Equation (1.1) the weights of the matrices $\mathbf{W}^k$ and offsets $\mathbf{b}^k$ are distributed according to (2.1), then it is possible to write (2.2) under the mean-field approximation as

$$q^{(\ell)} = \sigma_w^2 \int \mathcal{D}z \phi \left(\sqrt{q^{(\ell-1)}} z \right)^2 + \sigma_b^2 \tag{2.3}$$

where $\mathcal{D}z$ is the standard Gaussian measure.

It was proven in [75] that with certain activation functions, this measure converges to a fixed point $q^\star$ which allows to understand if all points are mapped to the origin or not.

Moreover, given two different inputs, $\mathbf{x}^{0,1}$ and $\mathbf{x}^{0,2}$, the following measure was introduced to relate their similarity while going through the different layers of the neural network, similar to a covariance,

$$q_{a,b}^{(\ell)} = \frac{1}{n_\ell} \sum_{i=1} \mathbf{h}_i^{(\ell)}(\mathbf{x}^{0,a}) \mathbf{h}_i^{(\ell)}(\mathbf{x}^{0,b}) \quad \text{with } a, b \in \{1, 2\}. \tag{2.4}$$

which converges under the mean-field approximation to the relation

$$q_{1,2}^{(\ell)} = \sigma_w^2 \int \mathcal{D}z_1 \mathcal{D}z_2 \phi(u_1) \phi(u_2) + \sigma_b^2 =: \mathcal{C}(c_{12}^{(\ell-1)}, q_{11}^{(\ell-1)}, q_{22}^{(\ell-1)} | \sigma_w, \sigma_b), \tag{2.5}$$

$$u_1 = \sqrt{q_{11}^{(\ell-1)}} z_1, \quad u_2 = \sqrt{q_{22}^{(\ell-1)}} \left[c_{12}^{(\ell-1)} z_1 + \sqrt{1 - (c_{12}^{(\ell-1)})^2} z_2 \right], \tag{2.6}$$

where $c_{12}^{(\ell)} = q_{12}^{(\ell)} (q_{11}^{(\ell)} q_{22}^{(\ell)})^{-1/2}$, $\mathcal{D}z_1$ and $\mathcal{D}z_2$ are independent standard Gaussian measures, and $\mathcal{C}$ is a newly introduced map.

The dynamics of two similar points both with energy $q^\star$, are thus described by the iterative relation for the correlation coefficient $c_{12}^{(\ell)}$

$$c_{12}^{(\ell)} = \frac{1}{q^\star} \mathcal{C}(c_{12}^{(\ell-1)}, q^\star, q^\star | \sigma_w, \sigma_b) \tag{2.7}$$

which has the fixed point $c^\star = 1$. The stability of this point is defined by the slope of the map at 1, which is defined by

$$\chi_1 = \left.\frac{\partial c_{12}^{(\ell)}}{\partial c_{12}^{(\ell-1)}}\right|_{c=1} = \sigma_w^2 \int \mathcal{D}z \left[\phi'(\sqrt{q^\star}z)\right]^2 . \tag{2.8}$$

If $\chi_1 < 1$ then the map $\mathcal{C}$ is stable at 1 and nearby points will converge to the same solution. However if $\chi_1 > 1$ the fixed point is unstable and two nearby points will behave as in a chaotic system. According to Poole et al. [76], the ability to disentangle manifolds is strictly related to the inputs in a neural network not converging to a unique solution, hence an expressive neural network should be such that $\chi_1 \geq 1$. This distinction divides the parameter space (σ_b, σ_w) with $\chi = 1$ denoted as the *edge of chaos*, see Figure 2.1 for an illustration when $\phi(\cdot) = tanh(\cdot)$[2]. Furthermore, Schoenholz et al. [88] associated the rate at which $q^{(\ell)}$ converges to $q^\star$, and $c_{12}^{(\ell)}$ converges to $c^\star$, to the depth over which information of the input may propagate and showed that this is maximised on the edge of chaos.

2.1.2 Gradient Flow

When training very deep neural networks, the gradients of the objective function with respect to the neural networks' parameters are computed via backpropagation [83]. As shown in [73], the gradients are prone to explode or vanish. Building on the results in [74, 88] studied the initialisation of the neural network, which allows the gradients in expectation neither to explode nor to vanish. In feed-forward neural networks the Jacobian $\mathbf{J} \in \mathbb{R}^{n,n}$ is given by

$$\mathbf{J} = \frac{\partial \mathbf{x}^{(L)}}{\partial \mathbf{x}^{(0)}} = \prod_{\ell=1}^{L} \mathbf{D}^{(\ell)}\mathbf{W}^{(\ell)} = \prod_{\ell=1}^{L} \mathbf{J}^{(\ell)} \tag{2.9}$$

where $\mathbf{D}^{(\ell)}$ is a diagonal matrix with entries relying on the activation function such that $\mathbf{D}_{ii}^{(\ell)} = \phi'(\mathbf{h}_i^{(\ell)})$.

The parameter χ_1 introduced above, was found in [74] to be related to the elements in the Jacobian by the following identity

$$\chi_1 = \frac{1}{N}\langle tr\left((\mathbf{DW})^\top \mathbf{DW}\right)\rangle. \tag{2.10}$$

[2] $tanh(x) = \frac{e^x - e^{-x}}{e^x + e^{-x}}$

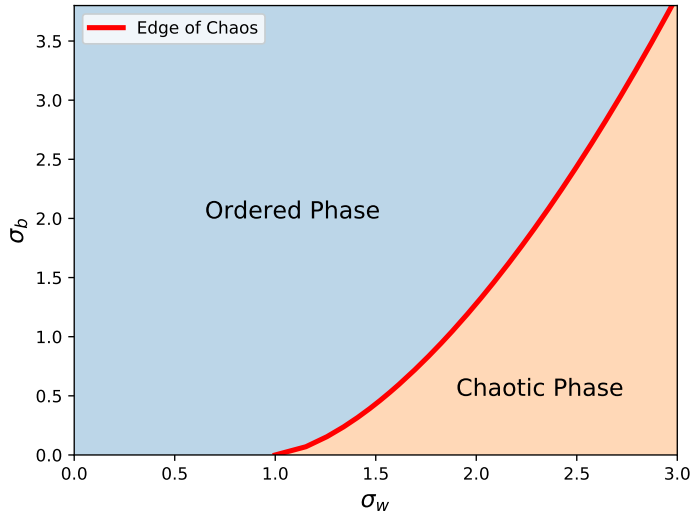


Figure 2.1: Illustration of the ordered-chaos transition when $\phi(h) = tanh(h)$ verifying the work in [88]. The critical line $\chi = 1$ defines the boundary between the two phases. In the chaotic regime $\chi > 1$ and gradients explode in expectation, while in the ordered regime $\chi < 1$ and gradients vanish in expectation.

It follows that χ_1 is the second moment of the distribution of squared singular values of the matrix $\mathbf{DW}$ and that the expected squared singular value of $\mathbf{J}$ is χ_1^L. Consequently, only for $\chi_1 = 1$ the gradients will not explode in expectation.

Modelling the weights of $\mathbf{W}^{(\ell)}$ to be either random Gaussian or random orthogonal, i.e. drawn from a uniform distribution over scaled orthogonal matrices obeying $\mathbf{W}^{(\ell)\top}\mathbf{W}^{(\ell)} = \sigma_w^2\mathbf{I}$, it is possible to analyse the spectrum of the Jacobian using tools from free probability theory [95]. This analysis proves that χ_1 is the expected squared singular value and also introduces an analysis for the variance, $\sigma^2_{J^\top J}$, of these squared singular values to show that choosing an initialisation with $\chi_1 = 1$ might not be sufficient to avoid the gradients exploding or vanishing. Although σ_w and σ_b might be tuned critically to have $\chi_1 = 1$, the variance $\sigma^2_{J^\top J}$ might increase linearly with the depth and thus the neural network could still have gradients exploding or vanishing.

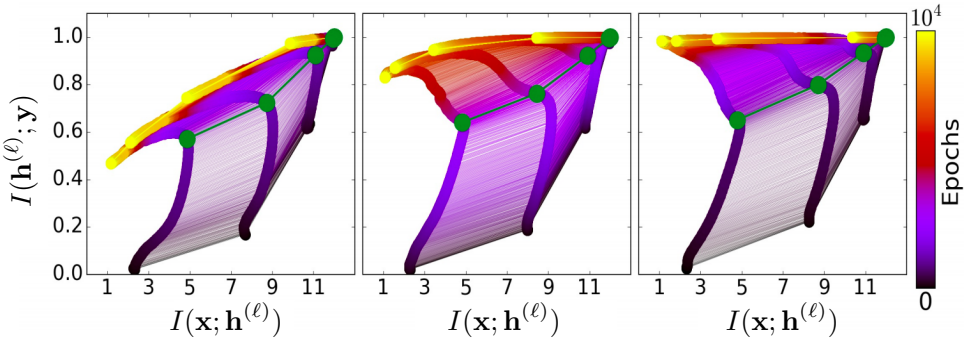


Figure 2.2: Figure reproduced from [91] illustrating how the mutual information between the input and the hidden layers, and between the hidden layers and the output evolve during training for different training dataset dimension (5% of the data on the left, 45% in the middle, and 85% on the right). The number of training epochs for each state is described by its colour. For more details see [91]

2.1.3 Mutual Information Flow

The work in [1] aimed at studying how information flows through a random neural network, and it considered using the mutual information measure from information theory rather than using a geometric interpretation as in [88]. The mutual information measure is used to quantify the amount of information that is carried through a noisy channel from a random variable $\mathbf{x} \in \mathcal{X}$ to a random variable $\mathbf{y} \in \mathcal{Y}$, and it is determined by the following equation

$$I(\mathbf{x};\mathbf{y}) = D_{KL}(P_{(\mathbf{x},\mathbf{y})}||P_{\mathbf{x}} \otimes P_{\mathbf{y}}) = \int_{\mathcal{X}} \int_{\mathcal{Y}} p_{\mathbf{x},\mathbf{y}}(\mathbf{x},\mathbf{y}) \log \left(\frac{p_{\mathbf{x},\mathbf{y}}(\mathbf{x},\mathbf{y})}{p_{\mathbf{x}}(\mathbf{x}) p_{\mathbf{y}}(\mathbf{y})} \right) \, d\mathbf{x} \, d\mathbf{y}. \tag{2.11}$$

The KL divergence, D_{KL}, measure a distance between two distributions and in this case it measures the distance between the joint distribution $P_{(\mathbf{x},\mathbf{y})}$ and the product of the marginal distributions $P_{\mathbf{x}} \otimes P_{\mathbf{y}}$. Since $P_{\mathbf{x}} \otimes P_{\mathbf{y}}$ would be the joint distribution if $\mathbf{x}$ and $\mathbf{y}$ were independent, the mutual information can be interpreted as the "error" of considering the two variables $\mathbf{x}$ and $\mathbf{y}$ as independent.

In [91] the use of the mutual information was used to describe the state of the training of neural networks by plotting the mutual information between an input $\mathbf{x}$ and the hidden layer $\mathbf{h}^{(\ell)}$, $I(\mathbf{x};\mathbf{h}^{(\ell)})$ against the mutual information between the hidden layer $\mathbf{h}^{(\ell)}$ and the output $\mathbf{y}$, $I(\mathbf{h}^{(\ell)};\mathbf{y})$. Thanks to examples such as

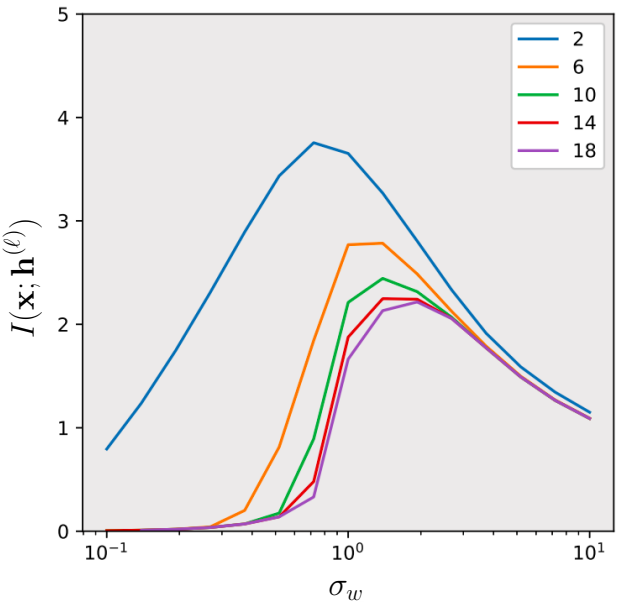


Figure 2.3: Figure reproduced from [1] depicting the convergence of the mutual information between the input $\mathbf{x}$ and the post-activation hidden layer $\mathbf{h}^{(\ell)}$ for a random neural network with $\sigma_b = 0$ and $\phi(\cdot) = Htanh(\cdot)$[3]when σ_w varies.

Figure 2.2, it was shown how during the initial phases of training the update of the parameters was maximising the information carried through the layers and in a later stage it was compressing the information. In Figure 2.2 the change of phase corresponds to the green line. However, the concept of compression in [91] raised different concerns because of how the mutual information was approximated [30, 85] and it was actually shown in [30] that the compression phase was actually describing a clustering of the hidden representations.

In this context, the work in [1] tried to understand if there is an initialisation that would maximise the mutual information and therefore result in a faster training. By plotting the mutual information approximation in [28] for different layers when the standard deviation of the weights in the initialisation (2.1) varies, it was

[3]

$$Htanh(x) = \begin{cases} 1 & \text{if } x > 1 \\ x & \text{if } |x| \leq 1 \\ -1 & \text{if } x < -1 \end{cases}$$

observed that there is a choice of parameters such that the mutual information between the input and the final layer is maximised. This behaviour is presented in Figure 2.3 where the mutual information is maximised for $\sigma_w \approx 1.4$ when $\sigma_b = 0$.

The mutual information estimation is a challenging task and its analytical formulation is most often not existent [69]. Therefore, the studies involving its estimation usually rely on an approximation that is computed with some sampling-procedures [23, 26, 28, 30, 52, 68, 77, 85]. These kinds of analyses suffer a strong dependency on the number of data samples [58], do not always agree [1], and do not give insights on how the different variables affect the mutual information. Therefore, the work in this thesis validates the observations in [1] from a more theoretical perspective giving an insight into how the initialisation parameters affect the mutual information propagation.

2.2 A Gaussian Lower Bound on the Mutual Information

In [1], it was observed experimentally that the mutual information planes advocated by [91] depend strongly on the parameters $(\sigma_w, \sigma_b, \phi(\cdot))$. The mathematical analysis here complements the observations in [1] by introducing an alternative lower bound of the mutual information.

We are interested in computing the mutual information $I(\mathbf{x}; \mathbf{h}^{(\ell)})$ between the input $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \sigma_x^2 \mathbf{I})$, as also modelled in [16], and the hidden layer $\mathbf{h}^{(\ell)}$ by conditioning on a realisation of the random weights $\mathcal{W}^{:\ell} = \{\mathbf{W}^{(i)}, \mathbf{b}^{(i)}\}_{i=1}^{\ell}$

$$I(\mathbf{x}; \mathbf{h}^{(\ell)}) = \mathbb{E}_{\mathcal{W}^{:\ell}} \left[I(\mathbf{x}; \mathbf{h}^{(\ell)} | \mathcal{W}^{:\ell}) \right]. \tag{2.12}$$

For this measure to be insightful, we need to consider a noise term $\mathbf{n}^{(\ell)}$ which is a Gaussian variable added before the activation function, as shown in Figure 2.4, and consequently perturbs the distribution of $\mathbf{h}^{(\ell)}$. This noise term is necessary because once the neural network map is sampled, $\mathbf{x} \mapsto \mathbf{h}^{(\ell)}$ is completely deterministic if there is no noise, and consequently, the mutual information $I(\mathbf{x}; \mathbf{h}^{(\ell)} | \mathcal{W}^{:\ell})$ between the hidden layers and the input is infinite [30]. Initially, this was solved by considering binning the values in the hidden variables [91], but the results were

found to be too dependent on the choices of the bins [85] The research in [30] and [26] showed that it is more appropriate to consider a random noise $\mathbf{n}^{(\ell)} \sim \mathcal{N}(0, \sigma_n^2)$ added at each layer.

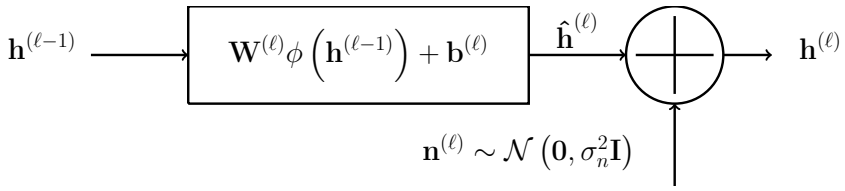


Figure 2.4: Signal Propagation in the perturbed neural network introduced in [30].

For two random variables with a generic distribution, the mutual information is not known explicitly and its approximation is a challenging task that has been attempted primarily with non-parametric models [28, 48, 66]. However, if two random variables $\mathbf{x}$ and $\mathbf{y}$ are jointly Gaussian with marginal covariances, $\mathbf{\Lambda}_x$ and $\mathbf{\Lambda}_y$, and joint covariance, $\mathbf{\Lambda}_{xy}$, the mutual information is

$$I^{\mathcal{G}}(\mathbf{x}; \mathbf{y}) = \frac{1}{2} \log \left(\frac{|\mathbf{\Lambda}_x| \, |\mathbf{\Lambda}_y|}{|\mathbf{\Lambda}_{xy}|} \right) \tag{2.13}$$

where $|\cdot|$ denotes the matrix determinant [40].

While the distribution of $\mathbf{h}^{(\ell)}$ is known to converge to a Gaussian distribution with large depth ℓ, in order to lower bound the mutual information throughout the layers we note that the mutual information for a general distribution is lower bounded by that of a Gaussian distribution.

In [40] it is shown that

Proposition 2.2.1 ([40]). *Let $g(\boldsymbol{x})$ be an n-dimensional Gaussian distribution , $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Lambda})$, with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Lambda}$. If $f(\boldsymbol{x})$ is an arbitrary distribution with the same mean and covariance matrix, then*

$$\mathbb{E}_f[-log(g(\boldsymbol{x}))] = \mathbb{E}_g[-log(g(\boldsymbol{x}))] \tag{2.14}$$

Proof.

$$-\mathbb{E}_f[log(g(\mathbf{x}))] = -\mathbb{E}_f \left[log \frac{1}{\sqrt{(2\pi)^n |\mathbf{\Lambda}|}} e^{-\frac{1}{2}(\mathbf{x}-\boldsymbol{\mu})^\top \mathbf{\Lambda}^{-1}(\mathbf{x}-\boldsymbol{\mu})} \right] \tag{2.15}$$

$$= \frac{1}{2} log(|2\pi\mathbf{\Lambda}|) + \frac{log(e)}{2}\mathbb{E}_f\left[(\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{\Lambda}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] \tag{2.16}$$

$$= \frac{1}{2} log(|2\pi\mathbf{\Lambda}|) + \frac{n\log(e)}{2} \tag{2.17}$$

$$= \frac{1}{2} log(|2\pi e\mathbf{\Lambda}|) \tag{2.18}$$

$$= -\mathbb{E}_g\left[log(g(\mathbf{x}))\right] \tag{2.19}$$

where in (2.16) we considered the fact that for any such function f

$$\mathbb{E}_f\left[(\mathbf{x}-\boldsymbol{\mu})^\top\mathbf{\Lambda}^{-1}(\mathbf{x}-\boldsymbol{\mu})\right] = \mathbb{E}_f\left[\sum_{i,j=1}^{n}(\mathbf{x}-\boldsymbol{\mu})_i^\top(\mathbf{\Lambda}^{-1})_{ij}(\mathbf{x}-\boldsymbol{\mu})_j\right] \tag{2.20}$$

$$= \sum_{i,j=1}^{n}\mathbb{E}_f\left[(\mathbf{x}-\boldsymbol{\mu})_i^\top(\mathbf{\Lambda}^{-1})_{ij}(\mathbf{x}-\boldsymbol{\mu})_j\right] \tag{2.21}$$

$$= \sum_{i,j=1}^{n}(\mathbf{\Lambda}^{-1})_{ij}\mathbb{E}_f\left[(\mathbf{x}-\boldsymbol{\mu})_i^\top(\mathbf{x}-\boldsymbol{\mu})_j\right] \tag{2.22}$$

$$= \sum_{i,j=1}^{n}(\mathbf{\Lambda}^{-1})_{ij}\mathbf{\Lambda}_{i,j} = \sum_{i,j=1}^{n}(\mathbf{\Lambda}^{-1})_{ij}\mathbf{\Lambda}_{j,i} \tag{2.23}$$

$$= tr(\mathbf{\Lambda}^{-1}\mathbf{\Lambda}) = tr(\mathbf{I}) = n \tag{2.24}$$

and in (2.16) we used the property that for any scalar a and a matrix $\mathbf{X}$ the following holds $|a\mathbf{X}| = a^n|\mathbf{X}|$.

In equation (2.18) we considered that, since a Gaussian distributed vector $\mathbf{y} \sim \mathcal{N}(\mathbf{0}, \mathbf{I})$ with probability density function $g_{\mathbf{y}}(\cdot)$ is such that

$$\mathbb{E}_g[-\log(g(\mathbf{y}))] = -\int_{\mathbb{R}^n}\frac{1}{(2\pi)^{n/2}}e^{-\frac{1}{2}\mathbf{y}^\top\mathbf{y}}\log\left(\frac{1}{(2\pi)^{n/2}}e^{-\frac{1}{2}\mathbf{y}^\top\mathbf{y}}\right)d\mathbf{y} \tag{2.25}$$

$$= \frac{n}{2}\log(2\pi) + \frac{\log(e)}{2}\int_{\mathbb{R}^n}\frac{1}{(2\pi)^{n/2}}e^{-\frac{1}{2}\mathbf{y}^\top\mathbf{y}}\mathbf{y}^\top\mathbf{y}d\mathbf{y} \tag{2.26}$$

$$= \frac{n}{2}\log(2\pi) + \frac{n\log(e)}{2} = \frac{n}{2}\log(2\pi e) \tag{2.27}$$

and $g(\mathbf{x}) = \frac{1}{\mathbf{\Lambda}^{-1/2}}g_{\mathbf{y}}(\mathbf{\Lambda}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}))$, then the expected logarithm is

$$\mathbb{E}_g\left[\log(g(\mathbf{x}))\right] = \int_{\mathbb{R}^n} g(\mathbf{x})\log(g(\mathbf{x}))d\mathbf{x} \tag{2.28}$$

$$= \int_{\mathbb{R}^n}\frac{1}{|\mathbf{\Lambda}^{-1/2}|}g_{\mathbf{y}}(\mathbf{\Lambda}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}))\log\frac{1}{|\Lambda_{\mathbf{X}}^{-1/2}|}g_{\mathbf{y}}(\mathbf{\Lambda}^{-1/2}(\mathbf{x}-\boldsymbol{\mu}))d\mathbf{x} \tag{2.29}$$

$$= \int_{\mathbb{R}^n} g_{\mathbf{y}}(\mathbf{y}) \log \frac{1}{|\mathbf{\Lambda}^{-1/2}|} g_{\mathbf{y}}(\mathbf{y}) d\mathbf{y} \tag{2.30}$$

$$= \int_{\mathbb{R}^n} g_{\mathbf{y}}(\mathbf{y}) \log g_{\mathbf{y}}(\mathbf{y}) d\mathbf{y} - \log |\mathbf{\Lambda}^{-1/2}| \int_{\mathbb{R}^n} g_{\mathbf{y}}(\mathbf{y}) d\mathbf{y} \tag{2.31}$$

$$= \mathbb{E}_{\mathbf{y}}[\log(\mathbf{y})] - \log |\mathbf{\Lambda}^{-1/2}| \tag{2.32}$$

$$= \frac{n}{2} \log(2\pi e) + \frac{1}{2} \log(|\mathbf{\Lambda}|) = \frac{1}{2} \log(|2\pi e \mathbf{\Lambda}|) \tag{2.33}$$

where in (2.33) we used again the property that for any scalar a and a matrix $\mathbf{X}$ the following holds $|a\mathbf{X}| = a^n |\mathbf{X}|$. □

This allows to prove the following lower bound on n-dimensional random variables.

Proposition 2.2.2 ([99, Proposition 1]). *Let $g_{\boldsymbol{xy}}(\boldsymbol{x}, \boldsymbol{y})$ be an n-dimensional Gaussian distribution on the vector $[\boldsymbol{x}^\top, \boldsymbol{y}^\top]^\top$, $\mathcal{N}(\boldsymbol{\mu}, \mathbf{\Lambda})$, with mean $\boldsymbol{\mu}$ and covariance matrix $\mathbf{\Lambda}$. If $f_{\boldsymbol{xy}}(\boldsymbol{x}, \boldsymbol{y})$ is an arbitrary distribution on the vector $[\boldsymbol{x}^\top, \boldsymbol{y}^\top]^\top$ with the same mean and covariance matrix, and with the marginal distribution $f_{\boldsymbol{x}}(\boldsymbol{x})$ Gaussian, then*

$$I_f(\boldsymbol{x}; \boldsymbol{y}) \geq I_g(\boldsymbol{x}; \boldsymbol{y}) \tag{2.34}$$

Proof.

$$\begin{aligned} I_f(\mathbf{x};\mathbf{y}) - I_g(\mathbf{x};\mathbf{y}) &= \int_{\mathcal{X},\mathcal{Y}} f_{\mathbf{xy}}(\mathbf{x},\mathbf{y}) \log \left(\frac{f_{\mathbf{xy}}(\mathbf{x},\mathbf{y})}{f_{\mathbf{x}}(\mathbf{x}) f_{\mathbf{y}}(\mathbf{y})} \right) d\mathbf{x} d\mathbf{y} \\ &\quad - \int_{\mathcal{X},\mathcal{Y}} g_{\mathbf{xy}}(\mathbf{x},\mathbf{y}) \log \left(\frac{g_{\mathbf{xy}}(\mathbf{x},\mathbf{y})}{g_{\mathbf{x}}(\mathbf{x}) g_{\mathbf{y}}(\mathbf{y})} \right) d\mathbf{x} d\mathbf{y} \qquad (2.35) \\ &= -\int_{\mathcal{X}} f_{\mathbf{x}}(\mathbf{x}) \log \left(f_{\mathbf{x}}(\mathbf{x}) \right) d\mathbf{x} - \int_{\mathcal{Y}} f_{\mathbf{y}}(\mathbf{y}) \log \left(f_{\mathbf{y}}(\mathbf{y}) \right) d\mathbf{y} \\ &+ \int_{\mathcal{X},\mathcal{Y}} f_{\mathbf{xy}}(\mathbf{x},\mathbf{y}) \log \left(f_{\mathbf{xy}}(\mathbf{x},\mathbf{y}) \right) d\mathbf{x} d\mathbf{y} + \int_{\mathcal{X}} g_{\mathbf{x}}(\mathbf{x}) \log \left(g_{\mathbf{x}}(\mathbf{x}) \right) d\mathbf{x} \\ &+ \int_{\mathcal{Y}} g_{\mathbf{y}}(\mathbf{y}) \log \left(g_{\mathbf{y}}(\mathbf{y}) \right) d\mathbf{y} - \int_{\mathcal{X},\mathcal{Y}} g_{\mathbf{xy}}(\mathbf{x},\mathbf{y}) \log \left(g_{\mathbf{xy}}(\mathbf{x},\mathbf{y}) \right) d\mathbf{x} d\mathbf{y} \qquad (2.36) \\ &\overset{f_{\mathbf{x}}=g_{\mathbf{x}}}{=} -\int_{\mathcal{Y}} f_{\mathbf{y}}(\mathbf{y}) \log \left(f_{\mathbf{y}}(\mathbf{y}) \right) d\mathbf{y} + \int_{\mathcal{X},\mathcal{Y}} f_{\mathbf{xy}}(\mathbf{x},\mathbf{y}) \log \left(f_{\mathbf{xy}}(\mathbf{x},\mathbf{y}) \right) d\mathbf{x} d\mathbf{y} \end{aligned}$$

$$+\int_{\mathcal{Y}} g_{\mathbf{y}}(\mathbf{y}) \log\left(g_{\mathbf{y}}(\mathbf{y})\right) d\mathbf{y} - \int_{\mathcal{X},\mathcal{Y}} g_{\mathbf{xy}}(\mathbf{x},\mathbf{y}) \log\left(g_{\mathbf{xy}}(\mathbf{x},\mathbf{y})\right) d\mathbf{x}d\mathbf{y} \tag{2.37}$$

$$\overset{Prop.2.2.1}{=} \int_{\mathcal{Y}} f_{\mathbf{y}}(\mathbf{y}) \log\left(\frac{g_{\mathbf{y}}(\mathbf{y})}{f_{\mathbf{y}}(\mathbf{y})}\right) d\mathbf{y} - \int_{\mathcal{X},\mathcal{Y}} f_{\mathbf{xy}}(\mathbf{x},\mathbf{y}) \log\left(\frac{g_{\mathbf{xy}}(\mathbf{x},\mathbf{y})}{f_{\mathbf{xy}}(\mathbf{x},\mathbf{y})}\right) d\mathbf{x}d\mathbf{y} \tag{2.38}$$

$$= \int_{\mathcal{X},\mathcal{Y}} f_{\mathbf{xy}}(\mathbf{x},\mathbf{y}) \log\left(\frac{g_{\mathbf{y}}(\mathbf{y})}{f_{\mathbf{y}}(\mathbf{y})}\right) d\mathbf{x}d\mathbf{y} - \int_{\mathcal{X},\mathcal{Y}} f_{\mathbf{xy}}(\mathbf{x},\mathbf{y}) \log\left(\frac{g_{\mathbf{xy}}(\mathbf{x},\mathbf{y})}{f_{\mathbf{xy}}(\mathbf{x},\mathbf{y})}\right) d\mathbf{x}d\mathbf{y} \tag{2.39}$$

$$= \int_{\mathcal{X},\mathcal{Y}} f_{\mathbf{xy}}(\mathbf{x},\mathbf{y}) \log\left(\frac{g_{\mathbf{y}}(\mathbf{y})}{f_{\mathbf{y}}(\mathbf{y})} \frac{f_{\mathbf{xy}}(\mathbf{x},\mathbf{y})}{g_{\mathbf{xy}}(\mathbf{x},\mathbf{y})}\right) d\mathbf{x}d\mathbf{y} \tag{2.40}$$

$$= \int_{\mathcal{Y}} f(\mathbf{y}) \int_{\mathcal{X}} f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y}) \log\left(\frac{f_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})}{g_{\mathbf{x}|\mathbf{y}}(\mathbf{x}|\mathbf{y})}\right) d\mathbf{x}d\mathbf{y} \tag{2.41}$$

$$= \int_{\mathcal{Y}} f_{\mathbf{y}}(\mathbf{y}) D_{KL}(f_{\mathbf{x}|\mathbf{y}} || g_{\mathbf{x}|\mathbf{y}}) d\mathbf{y} \geq 0 \tag{2.42}$$

where the third equality is due to $f_{\mathbf{x}}$ being Gaussian, hence $f_{\mathbf{x}} = g_{\mathbf{x}}$; the fourth is due to Proposition 2.2.1. □

The analysis herein relies on the first two moments of $\left[\mathbf{x}^\top,\ \mathbf{h}^{(\ell)\top}\right]^T$ which for neural networks drawn according to (2.1) with $\mathbf{x} \sim \mathcal{N}(\mathbf{0}, \sigma_x^2\mathbf{I})$ and $\{\mathbf{n}^{(\ell)} \sim \mathcal{N}(0, \sigma_n^2\mathbf{I})\}_{l=1}^{\ell}$are:

$$\mathbb{E}_{\mathbf{x},\{\mathbf{n}\}_{l=1}^{\ell}|\mathcal{W}:\ell} \begin{bmatrix} \mathbf{x} \\ \mathbf{h}^{(\ell)} \end{bmatrix} = \begin{bmatrix} \mathbf{0} \\ \boldsymbol{\mu} \end{bmatrix} \tag{2.43}$$

$$\boldsymbol{\Lambda}_{xh^{(\ell)}} := \mathbb{V}ar_{\mathbf{x},\{\mathbf{n}\}_{l=1}^{\ell}|\mathcal{W}:\ell} \begin{bmatrix} \mathbf{x} \\ \mathbf{h}^{(\ell)} \end{bmatrix} = \begin{bmatrix} \sigma_x^2\mathbf{I} & \boldsymbol{\Sigma}_{xh^{(\ell)}} \\ \boldsymbol{\Sigma}_{xh^{(\ell)}}^\top & \boldsymbol{\Lambda}_{h^{(\ell)}} \end{bmatrix}. \tag{2.44}$$

In order to compute $I^{\mathcal{G}}$ in (2.13), we reformulate the determinant of the block covariance matrix (2.44) using the decomposition [97]:

$$\begin{bmatrix} \mathbf{E} & \mathbf{F} \\ \mathbf{F}^\top & \mathbf{C} \end{bmatrix} = \begin{bmatrix} \mathbf{I} & \mathbf{0} \\ \mathbf{F}^\top\mathbf{E}^{-1} & \mathbf{I} \end{bmatrix} \begin{bmatrix} \mathbf{E} & \mathbf{0} \\ \mathbf{0} & \mathbf{C} - \mathbf{F}^\top\mathbf{E}^{-1}\mathbf{F} \end{bmatrix} \begin{bmatrix} \mathbf{I} & \mathbf{E}^{-1}\mathbf{F} \\ \mathbf{0}^\top & \mathbf{I} \end{bmatrix}. \tag{2.45}$$

Thus the determinant is

$$\begin{vmatrix} \boldsymbol{\Lambda}_x & \boldsymbol{\Lambda}_{xy} \\ \boldsymbol{\Lambda}_{xy}^\top & \boldsymbol{\Lambda}_{\boldsymbol{y}} \end{vmatrix} = |\boldsymbol{\Lambda}_x| \left|\boldsymbol{\Lambda}_y - \boldsymbol{\Lambda}_{xy}^\top \boldsymbol{\Lambda}_x^{-1} \boldsymbol{\Lambda}_{xy}\right|. \tag{2.46}$$

Incorporating (2.46) for the determinant of $\mathbf{\Lambda}_{xh^{(\ell)}}$ in (2.13) we define the following lower bound for the mutual information for a neural network conditional on a set of weights $\mathcal{W}^{:\ell}$ with Gaussian input

$$I(\mathbf{x},\mathbf{h}^{(\ell)}|\mathcal{W}^{:\ell}) \geq I^{\mathcal{G}}(\mathbf{x},\mathbf{h}^{(\ell)}) = \frac{1}{2}\log\left(\frac{|\mathbf{\Lambda}_{h^{(\ell)}}|}{\left|\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^{\top}\mathbf{\Sigma}_{xh^{(\ell)}}\right|}\right). \tag{2.47}$$

Thus, from (2.12) the mutual information of a neural network is bounded as follows

$$I(\mathbf{x},\mathbf{h}^{(\ell)}) \geq \frac{1}{2}\underbrace{\mathbb{E}_{\mathcal{W}^{:\ell}}\left[\log\left(|\mathbf{\Lambda}_{h^{(\ell)}}|\right)\right]}_{(E_1)} \quad \underbrace{-\mathbb{E}_{\mathcal{W}^{:\ell}}\left[\frac{1}{2}\log\left(\left|\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^{\top}\mathbf{\Sigma}_{xh^{(\ell)}}\right|\right)\right]}_{(E_2)}. \tag{2.48}$$

2.2.1 Numerical Evaluation of the Gaussian Lower Bound

The quantities (E_1) and (E_2) in (2.48) can be computed by sampling different realisations of the weights $\mathcal{W}^{:\ell}$ and by then averaging over the log-determinant of the matrices $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^{\top}\mathbf{\Sigma}_{xh^{(\ell)}}$. Given the Cholesky factorisation of a matrix $\mathbf{A} = \mathbf{L}\mathbf{L}^{\top}$, the log-determinant of the factorised matrix can be computed as follows

$$\log|\mathbf{A}| = \log|\mathbf{L}||\mathbf{L}^{\top}| = 2\log|\mathbf{L}| = 2\log\left(\prod_{j=1}^{n}\mathbf{L}_{ii}\right) = 2\sum_{i=1}^{n}\log(\mathbf{L}_{ii}). \tag{2.49}$$

Here we compute these estimates by sampling 10^5 different inputs $\mathbf{x}$ and noises $\mathbf{n}$ with $\sigma_x = 1$ and $\sigma_n = 0.1$. Figure 2.5 shows how the Gaussian lower bounds changes as a function of σ_w for neural networks with square weight matrices of size $n \times n$ with $n = 90$, with the *tanh* activation function, and with σ_b chosen such that (σ_w, σ_b) lies on the EoC.

These numerical experiments show how the lower bound on the mutual information between the input and the hidden layers decreases at each layer and is maximised for a value of (σ_w, σ_b) close to $(1, 0)$, as advocated in [75].

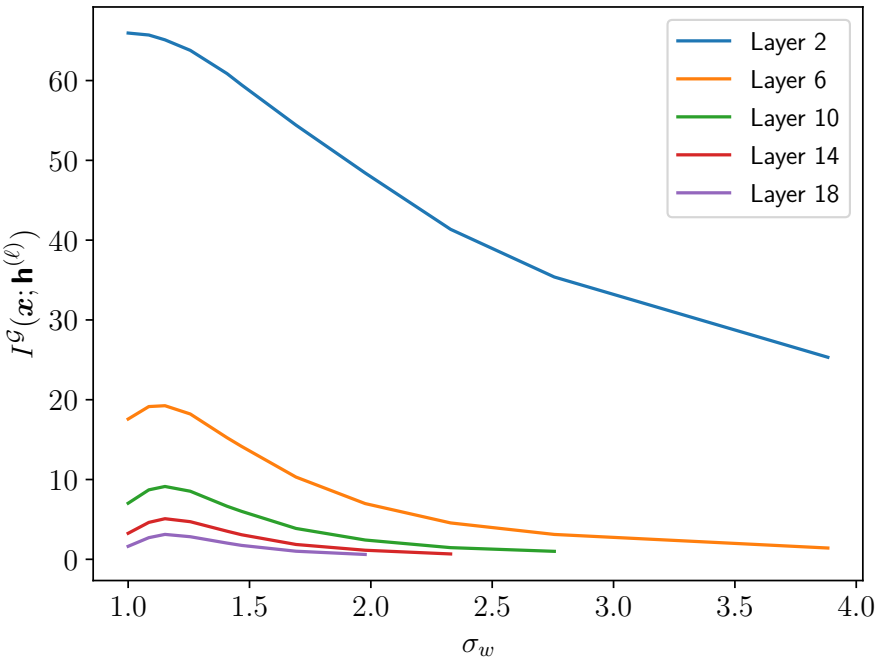


Figure 2.5: Numerical approximation by sampling of the Gaussian lower bound in (2.48) for the conditional mutual information of a feed-forward neural network when σ_w varies along the edge of chaos with $n = 90$ and $\phi(\cdot) = tanh(\cdot)$.

2.3 An Analytic Gaussian Lower Bound

The numerical computation of (E_1) and (E_2) based on sampling the covariance matrix is computationally expensive for large layer width n and depth L. Moreover, it less directly shows how this observed mutual information lower bound compares with the mean field analysis in [75] and the corresponding edge of chaos analysis. For easier computation and to better link these analyses, we approximate the matrices $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$ based on the mean field assumption of [75].

2.3.1 Mean Field Approximation

By the mean field analysis in [75], the hidden layers are normally distributed in the large limit, i.e. $n \gg 1$, with the mean and variance given in (2.43) and (2.44) respectively. In order to compute (2.48), we model the factors $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$ respectively by their expectation over the weights $\mathcal{W}^{:\ell}$.

2.3.1.1 Expectation of $\mathbf{\Lambda}_{h^{(\ell)}}$

The expectation of $\mathbf{\Lambda}_{h^{(\ell)}}$ in (2.48) was studied in [74, 75] and was shown that for large neural network width $n \gg 1$

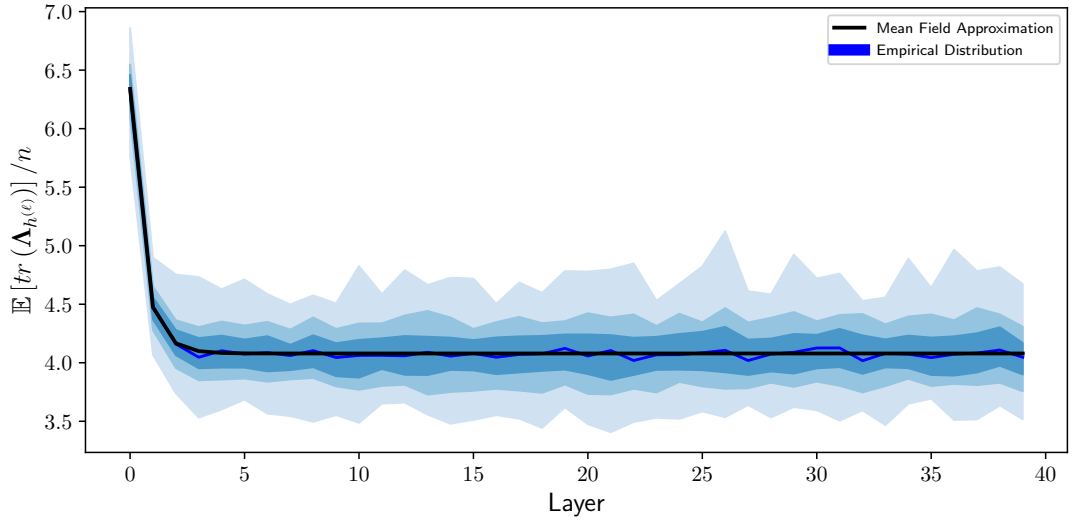
$$\mathbb{E}_{\mathcal{W}^{:\ell}}\left[\mathbf{\Lambda}_{h^{(\ell)}}\right] = q^{(\ell)}\mathbf{I}, \tag{2.50}$$

where $q^{(\ell)}$ corresponds to the variance of the hidden layer signal and is defined recursively via

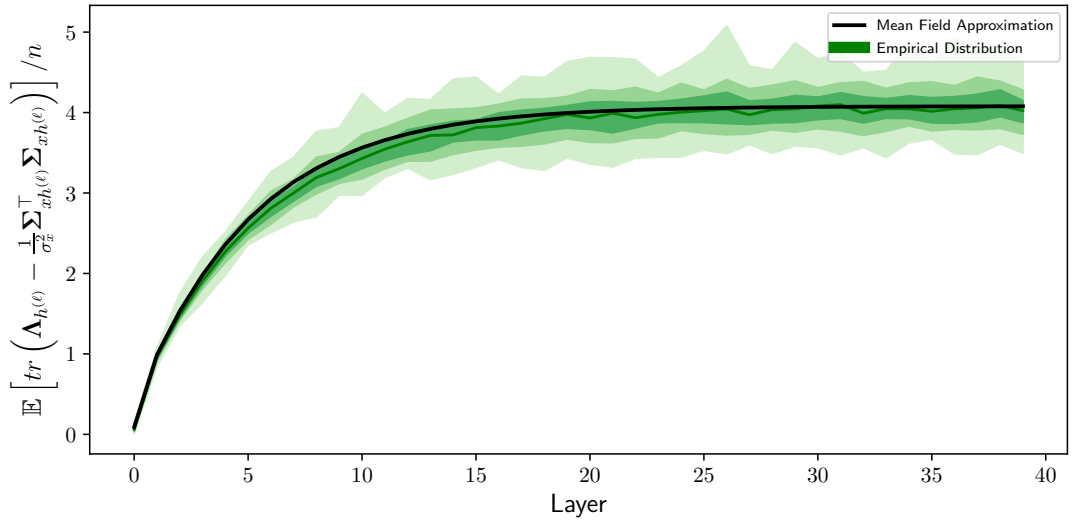
$$q^{(\ell)} = \sigma_w^2 \int \phi(\sqrt{q^{(\ell-1)}}z)\mathcal{D}z + \sigma_b^2 + \sigma_n^2 \tag{2.51}$$

$$q^{(1)} = \sigma_w^2\sigma_x^2 + \sigma_b^2 + \sigma_n^2. \tag{2.52}$$

In Figure 2.6a we demonstrate the validity of this approximation by plotting $q^{(\ell)}$ as the solid black line along with the empirical distributions obtained by generating 100 realisations of (1.1) for $(\sigma_w, \sigma_b) = (2.5, 0.3)$, $\phi(\cdot) = tanh(\cdot)$ and $n = 50$ and showing the distribution over 10^4 randomly drawn inputs $\mathbf{x}$. Note the good agreement of $q^{(\ell)}$ and the empirical values, as well as the limiting value for large depth, denoted q^* in [88]. The striking agreement is notable given the relatively



(a)



(b)

Figure 2.6: In (a) and (b) we compare mean field approximation of the mean value on the diagonal of the matrices $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}^{\top}_{xh^{(\ell)}}\mathbf{\Sigma}_{xh^{(\ell)}}$ respectively with the empirical distribution in function of the layers. The empirical distribution is obtained by considering 100 different set of weights $\mathcal{W}^{:\ell}$ for the neural network when $n = 50$ and with 10^5 different random inputs. The shading of the distribution corresponds to the 50%, 80%, and 100% confidence intervals in transparency order.

small neural network width $n = 50$ and the mean field approximation following from the $n \to \infty$ limit.

2.3.1.2 Expectation of $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$

To compute the expectation of the argument in the logarithm in (E_2) from (2.48), we can make use of (2.50) and then compute $\mathbb{E}_{\mathcal{W}^{:\ell}}\left[\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}\right]$. Since

$$\mathbb{E}_{\mathcal{W}^{:\ell}}\left[\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}\right]_{ij} = \mathbb{E}_{\mathcal{W}^{:\ell}}\left[\sum_k \mathbb{E}\left[\mathbf{x}_k \mathbf{h}_i^{(\ell)}\right] \mathbb{E}\left[\mathbf{x}_k \mathbf{h}_j^{(\ell)}\right]\right] \tag{2.53}$$

and as the weights in $\mathbf{W}^{(\ell)}$ are independent, the expected covariance matrix is diagonal with values

$$\mathbb{E}_{\mathcal{W}^{:\ell}}\left[\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}\right]_{ii} = \mathbb{E}_{\mathcal{W}^{:\ell}}\left[\sum_k \mathbb{E}\left[\mathbf{x}_k \mathbf{h}_i^{(\ell)}\right]^2\right]. \tag{2.54}$$

With the mean field analysis, we consider any elements i and j of respectively the input and the hidden layer to be jointly normally distributed according to

$$\begin{bmatrix} \mathbf{x}_i \\ \mathbf{h}_j^{(\ell)} \end{bmatrix} \sim \mathcal{N}\left(\begin{bmatrix} 0 \\ 0 \end{bmatrix}, \begin{bmatrix} \sigma_x^2 & \rho^{(\ell)}\sigma_x\sqrt{q^{(\ell)}} \\ \rho^{(\ell)}\sigma_x\sqrt{q^{(\ell)}} & q^{(\ell)} \end{bmatrix}\right) \tag{2.55}$$

with $\rho^{(\ell)}$ being the correlation coefficient. The correlation is given at each layer by solving

$$\rho^{(\ell)}\sigma_x\sqrt{q^{(\ell)}} = \underbrace{\int\int \sigma_x z_1 \frac{\sigma_w}{\sqrt{n}} \phi\left(\sqrt{q^{(\ell)}}\left(\rho^{(\ell-1)} z_1 + \sqrt{1-\rho^{(\ell-1)2}} z_2\right)\right) \mathcal{D}z_1 \mathcal{D}z_2}_{(E_3)} \tag{2.56}$$

where $\mathcal{D}z_1$ and $\mathcal{D}z_2$ are standard Gaussian measures.

The expected squared covariance is then given by

$$\mathbb{E}_{\mathcal{W}^{:\ell}}\left[\sum_k \mathbb{E}\left[\mathbf{x}_k \mathbf{h}_i^{(\ell)}\right]^2\right] = n(E_3)^2 \mathbf{I} \tag{2.57}$$

and consequently

$$\mathbb{E}\left[\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}\right] = (q^{(\ell)} - n(E_3)^2/\sigma_x^2)\mathbf{I} = q_c^{(\ell)}\mathbf{I}. \tag{2.58}$$

The validity of this approximation is shown in Figure 2.6b where as before we note the excellent agreement of the mean field limit (2.58) corresponding to width $n \to \infty$ and the observed distribution for (1.1) with width $n = 50$.

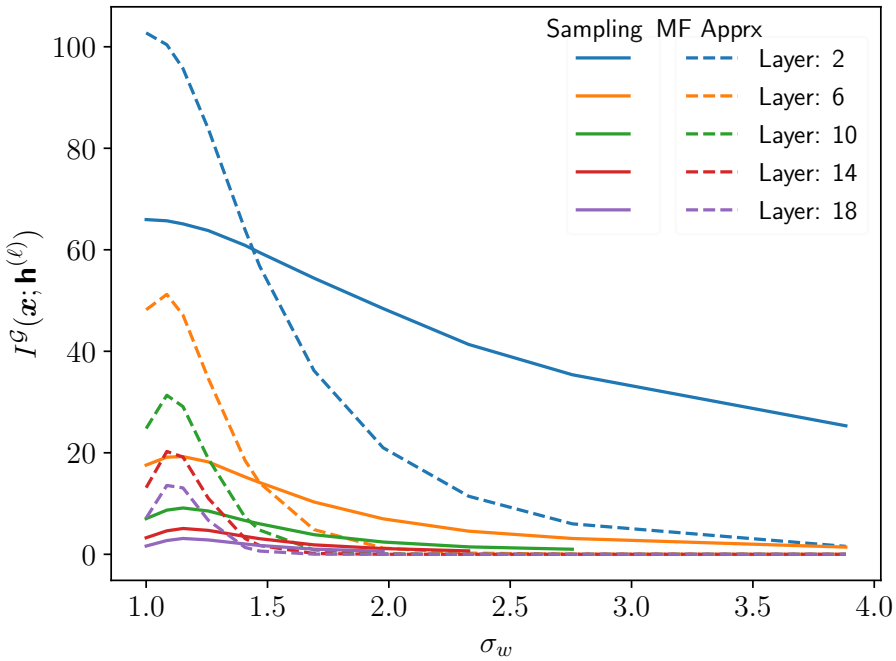


Figure 2.7: Comparison of the mean field approximation of the Gaussian lower bound $I^{\mathcal{G}}(\mathbf{x}, \mathbf{h}^{(\ell)})$ in (2.47) and its sampled value for a neural network with $n = 90$ and $tanh(\cdot)$ activation function when the variance σ_w is changed. Here σ_b is chosen so that (σ_w, σ_b) satisfies the EoC condition.

2.3.1.3 Lower Bound Approximation

Since the matrices $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^{\top}\mathbf{\Sigma}_{xh^{(\ell)}}$ are approximated as multiples of the identity matrix, their log-determinant can be easily computed; in particular

Proposition 2.3.1 ([99, Proposition 2]). *Under the mean field approximation the mutual information has the following lower bound*

$$I(\boldsymbol{X}, \boldsymbol{h}^{(\ell)}) \geq \frac{n}{2} \log \left(\frac{q^{(\ell)}}{q_c^{(\ell)}} \right) \tag{2.59}$$

Figure 2.7 compares the mean field approximation from Prop. 2.3.1 with the direct sampling approach as described in Figure 2.5 along the edge of chaos. We observe that the mean-field approximation is consistently overestimating the lower bound for small σ_w and underestimating for large σ_w. However, it seems to be

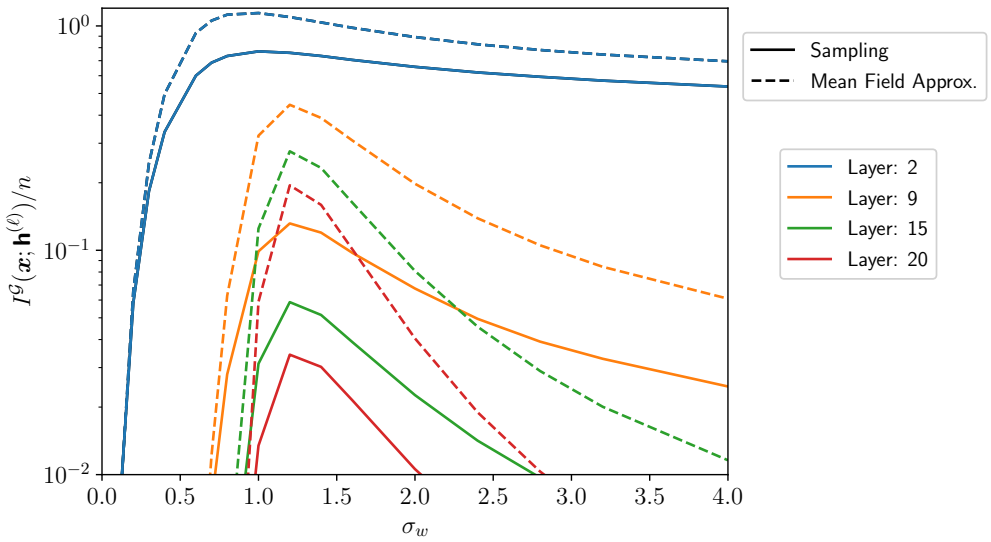
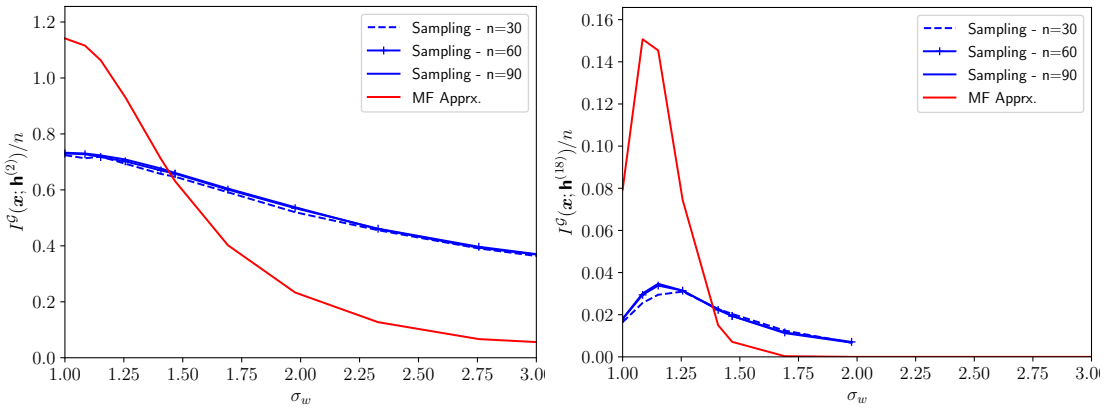


Figure 2.8: Comparison between the empirical lower bound and the one obtained with mean field approximation for a neural network with $n = 1000$ when the σ_w is changed on the x axis and σ_b is kept fixed to 0 with $\sigma_n = 0.1$.



(a) Mutual Information at Layer 2

(b) Mutual Information at Layer 18

Figure 2.9: Comparison of the sampled Gaussian lower bound $I^{\mathcal{G}}(\mathbf{x}, \mathbf{h}^{(\ell)})$ and the analytic approximation in (2.59) for $n = 30, 60$, and 90 for layers 2 (a) and 18 (b).

qualitatively reproducing the behaviour by predicting where the maximisers are. Figure 2.8 supports the same idea when σ_w varies with $\sigma_b = 0$ by showing how the mean-field approximation is consistently over-estimating the mutual information lower bound whilst being qualitatively accurate. Figure 2.9 illustrates the sampling and mean field calculations for varying widths $n = 30, 60, 90$ as well as at layers

$\ell = 1$ and 17. These results show how the inaccuracy cannot be due to the dimensionality of the problem itself and is most probably due to the assumption entailed by the mean-field approximation that the covariance matrices are diagonal.

2.4 Comparison of Mutual Information Approximation

The work in [1] showed that among different mutual information approximations as [48] and [66], the replica method in [28] is the most consistent measure with the arguments in [85], as it models with the decrease in mutual information for large enough variance σ_w^2 the loss in expressivity of the neural network due to the saturation of the *tanh* activation function. Figure 2.10 compares the replica method to the mean field lower bound for neural networks where the bias is null $\sigma_b = 0$, the noise has a variance $\sigma_n^2 = 10^{-5}$, and the input has i.i.d. normal elements, since the replica method is applicable only on these neural networks. The results show that the replica method and the approximated lower bound are both maximised for a value of the standard deviation σ_w thus supporting the saturation argument in [85]. The inconsistency for low standard deviations σ_w is due to the mean-field approximation overestimating the mutual information lower bound as shown in Figure 2.8.

Finally, in [1] the analysis with the replica method suggested that the mutual information information converges to a non-trivial limit as the depth increases. Figure 2.11 shows a consistent behaviour where the mean field lower bound also converges to a maximum for $\sigma_w = 1$ as the depth increases, with $\sigma_b = 0$ and $\phi(\cdot) = tanh(\cdot)$. Since $(\sigma_w, \sigma_b) = (1, 0)$ is on the edge of chaos for the *tanh* activation function, these calculations suggests that some initialisations on the edge of chaos, which considers a large depth limit, might be preferable both for optimisation [74] and mutual information. Specifically, the experiments reported in Figure 2.12 support that this double optimality holds only when $\sigma_b = 0$.

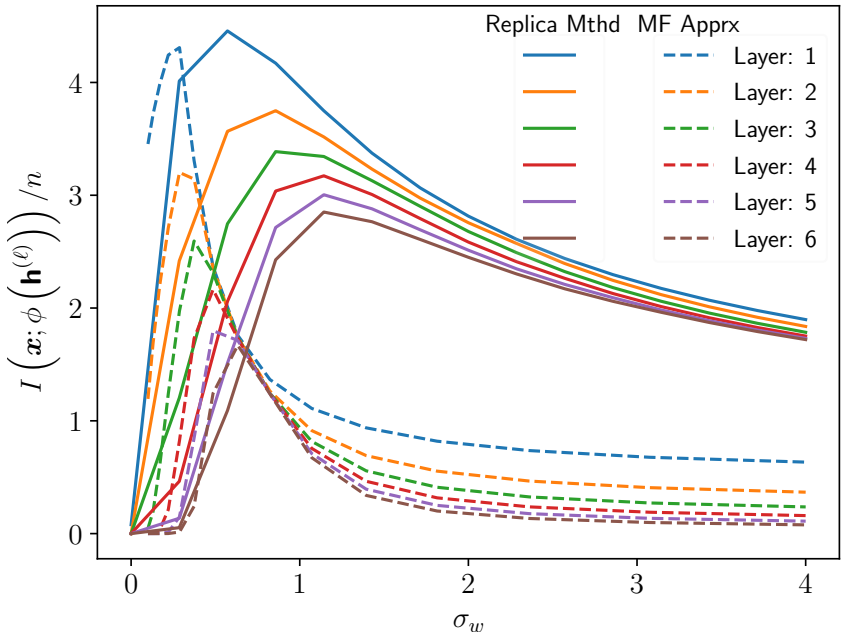


Figure 2.10: Comparison between the approximation of $I(\mathbf{x}, \mathbf{h}^{(\ell)})/n$ with the replica method (Replica Mthd) in [28] and the lower bound obtained with mean field approximation (MF Apprx) for a neural network with $n = 1000$ when the σ_w is changed on the x axis and σ_b is kept fixed. The results relative to this graph consider the post-activation hidden layer rather than the pre-activation hidden layer.

2.5 Considerations on the Mean-Field Approximation

We have presented a lower bound of the mutual information for feed-forward neural networks and we derived an approximation with the mean field theory which numerical experiments showed to have a qualitative representation of the original lower bound. The analytic lower bound approximation allows direct investigation of how the mutual information of neural networks change for different initialisation parameters $(\sigma_w, \sigma_b, \phi(\cdot))$. In particular, we observe that with $\phi(\cdot) = tanh(\cdot)$ activation function, the mutual information is maximised for σ_w on the edge of chaos when $\sigma_b = 0$ where the networks behave approximately linearly. In Chapter 3 we are going to study how this approximation can be improved by considering the eigen-distribution of the matrices $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^{\top}\mathbf{\Sigma}_{xh^{(\ell)}}$.

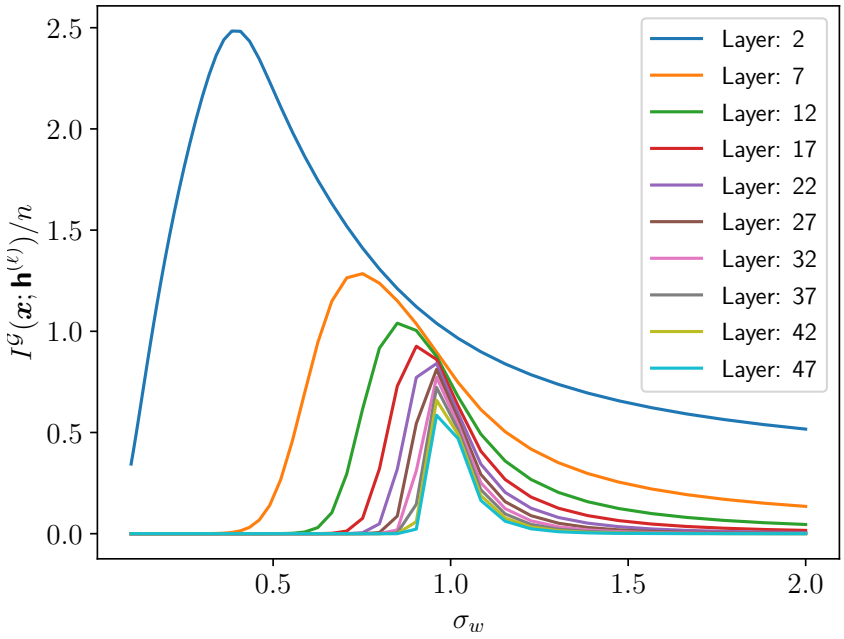


Figure 2.11: Convergence of the lower bound $I^{\mathcal{G}}(\mathbf{x}, \mathbf{h}^{(\ell)})/n$ obtained with the mean field approximation when we change the variance σ_w and keep $\sigma_b = 0$ and $\sigma_n = 1e-4$.

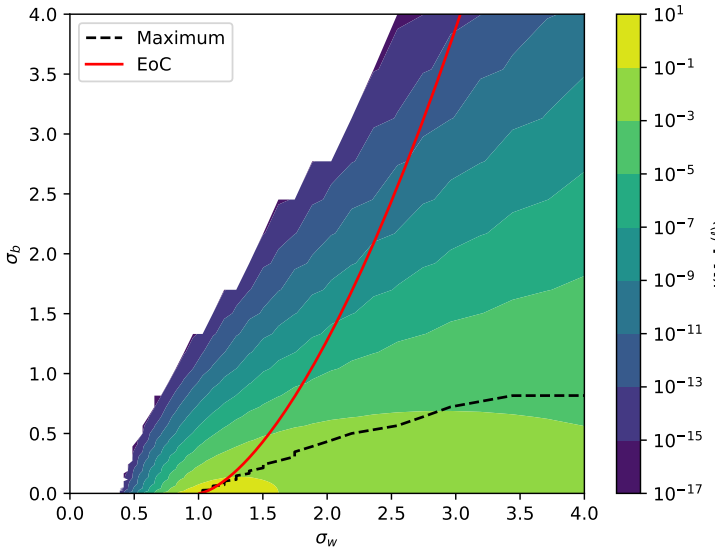


Figure 2.12: Lower bound $I^{\mathcal{G}}(\mathbf{x}, \mathbf{h}^{(28)})/n$ obtained with the mean field approximation when changing the variances (σ_w, σ_b). The dashed lines correspond to the maximum mutual information lower bound once σ_b is considered fixed.

Chapter 3

A Random Matrix Theory Approximation of the Mutual Information of Neural Networks at Initialisation

In the previous chapter, the study of the mutual information of neural networks was based on the assumption that the covariance matrices converged to their expected value. Here, we aim to improve the accuracy of the lower bound $I^{\mathcal{G}}(\mathbf{x}; \mathbf{h}^{(\ell)})$ in (2.47) by taking into consideration the full spectrum of the matrices $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$. If the eigen-distribution of the matrices $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$ are known it is possible to compute the mutual information exactly.

To compute the eigen-distribution of $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$, we will rely on results from Random Matrix Theory which are going to be presented in Section 3.1. This introductory section will also present the structure of Sections 3.2 and 3.3 where the eigen-distribution of $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$ are determined. In Section 3.4 the mutual information lower bound is approximated with the modelling of the eigen-distribution of $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$, showing an improvement in accuracy compared to Chapter 2. Some final remarks are included in Section 3.5

3.1 Introduction to Random Matrices

Random Matrix Theory focuses on the study of the properties of probability measures defined on spaces of matrices [47]. The questions that this field approaches arise from a number of different applications; among the many these include the condition number in linear algebra [21], the stability of high dimensional systems [57], and more applied problems such as the analysis of financial markets [45].

The work in random matrix theory relative to the functions defined on the eigenvalues is especially relevant in the mutual information lower bound estimation problem. In general, considering a function $F(\mathbf{A})$ that can be written as a power series on an n-dimensional matrix $\mathbf{A}$, the normalised trace of the function can be evaluated on the eigenvalues $\{\lambda_i\}_{i=1}^n$ of $\mathbf{A}$;

$$\frac{1}{n} trF(\mathbf{A}) = \frac{1}{n}\sum_{i=1}^{n} F(\lambda_i). \tag{3.1}$$

Therefore, for a large random matrix A, given its eigen-distribution $\rho_{\mathbf{A}}$, by the law of large numbers

$$\lim_{n\to\infty} \frac{1}{n}\sum_{i=1}^{n} F(\lambda_j) = \mathbb{E}_{\rho_{\mathbf{A}}}[F(\lambda)] \tag{3.2}$$

implying a concentration of measure; these scalar quantities do not fluctuate from sample to sample. In this chapter, this property is going to be applied to the log-determinant case. As a matter of fact, $log|\mathbf{A}| = tr(log(\mathbf{A}))$ and if the eigen-distribution $\rho_{\mathbf{A}}(\lambda)$ of matrix $\mathbf{A}$ is known, then

$$\lim_{n\to\infty} \log\left(|\mathbf{A}|\right)/n = \lim_{n\to\infty} \log\left(\prod_i \lambda_i\right)/n = \lim_{n\to\infty} \sum_i \log\left(\lambda_i\right)/n \tag{3.3}$$

$$= \int \log\left(x\right) d\rho_{\mathbf{A}}(x). \tag{3.4}$$

Therefore, by learning the eigen-spectrum of $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$ one is able to precisely compute their log-determinants whose difference defines the mutual information lower bound presented in (2.48) in Chapter 2.

In this introductory section, a baseline in Subsection 3.1.1 is presented where the eigen-distribution is approximated with known results for sample covariance

matrices. Then in Subsection 3.1.2 the Stieltjes transform is presented; its properties are central to the exact computation of the eigen-spectra relevant to the mutual information lower bound of a neural network. Finally, Section 3.1.3 details how the Stieltjes transform is implemented in Sections 3.2 and 3.3 to compute the eigen-spectra of $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$.

3.1.1 Marchenko-Pastur Law

The Marchenko-Pastur law is one of the most well-known results in Random Matrix Theory and it determines the eigen-spectrum of a sample covariance matrix when the dimensions of the sampled space and the number of samples go to infinity. Specifically, consider an $n \times m$ matrix $\mathbf{X}$ with entries $\mathbf{X}_{ij}$ that are i.id. real random variables such that $\mathbb{E}[\mathbf{X}_{ij}] = 0$ and $\mathbb{E}[\mathbf{X}_{ij}^2] = 1$, and denote by $\mathbf{M}$ the $n \times n$ matrix

$$\mathbf{M} = \frac{1}{m}\mathbf{X}\mathbf{X}^\top \in \mathbb{R}^{n \times n} \tag{3.5}$$

with $0 \leq \lambda_1 \leq \lambda_2 \leq \ldots \leq \lambda_n$ being the eigenvalues of $\mathbf{M}$.

Defining the random spectral distribution by

$$\mu_{\mathbf{M}}(x) := \frac{1}{n}\sum_{j=1}^{n} \mathbb{1}_{x > \lambda_j}, \tag{3.6}$$

with $\mathbb{1}$ being the indicator function, the Marchenko-Pastur law is defined in Theorem 3.1.1.

Theorem 3.1.1 (Marchenko-Pastur Law [54]). *For M and μ_M as defined above, when $n \to \infty$ and $m \to \infty$ such that $n/m \to \gamma \in (0, 1]$, $\mu_M \to \mu$ in expectation and almsot surely, where μ is the deterministic measure satisfying*

$$\frac{d\mu}{dx} = \begin{cases} \frac{1}{2\pi\gamma x}\sqrt{(a_+ - x)(x - a_-)} & \text{if } a_- \leq x \leq a_+ \\ 0 & \text{otherwise} \end{cases} \tag{3.7}$$

with $a_\pm = (1 \pm \sqrt{\gamma})^2$. When $\gamma > 1$, one needs to add $(1 - \gamma^{-1})\delta(x)$ to the right hand side of (3.7), where $\delta(x)$ denotes a Dirac delta-function/mass at the origin.

An example of how the Marchenko-Pastur law holds in practice is shown in Figure 3.1.

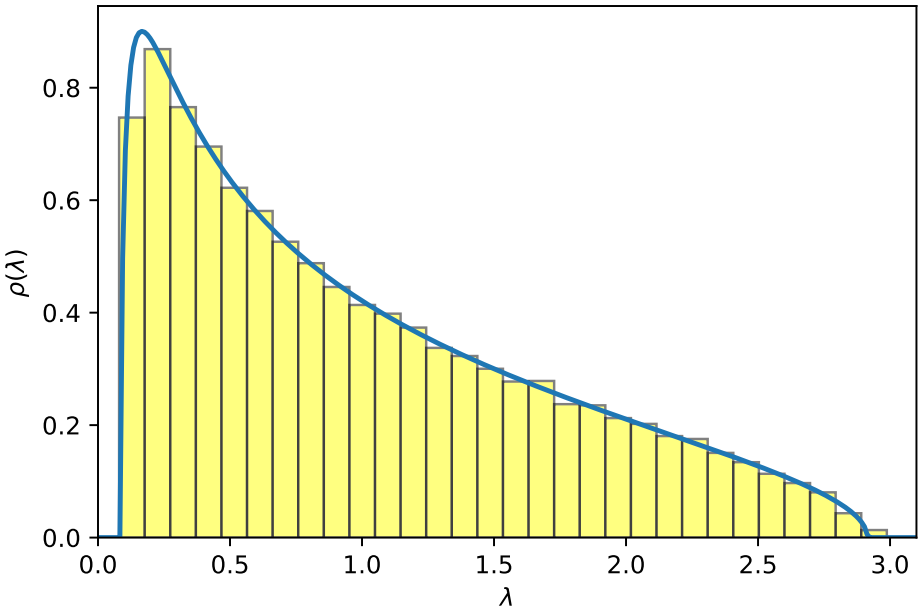


Figure 3.1: Example of how the Marchenko-Pastur distribution in Theorem 3.1.1 describes the eigen-distribution of a sample covariance in practice when $\gamma = 0.5$, $\sigma^2 = 1$, and $n = 100$.

3.1.1.1 Marchenko-Pastur Approximation for Mutual Information Computation

As a baseline for understanding the impact of using the eigen-spectra when computing the mutual information lower bound, this initial section approximates the eigen-spectra of $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$ with the Marchenko-Pastur law. In Figure 3.2 it is shown how the eigen-spectra of the empirical covariance matrices compares to the Marchenko-Pastur distribution when the covariance matrices were computed with n=200, $\sigma_w = 2.5$, $\sigma_x = 1$, and $\sigma_b = 0.3$.

For non unitary variance σ^2 of the elements $\mathbf{X}_{ij}$ in a Wishart matrix, the Marchenko-Pastur distribution is defined as

$$d\mu(x) = \frac{1}{2\pi\sigma^2}\frac{\sqrt{(\sigma^2 a_+ - x)(x - \sigma^2 a_-)}}{\gamma x}dx, \tag{3.8}$$

where $\gamma = n/m$ and $a_\pm = \sigma^2(1 \pm \sqrt{\gamma})^2$.

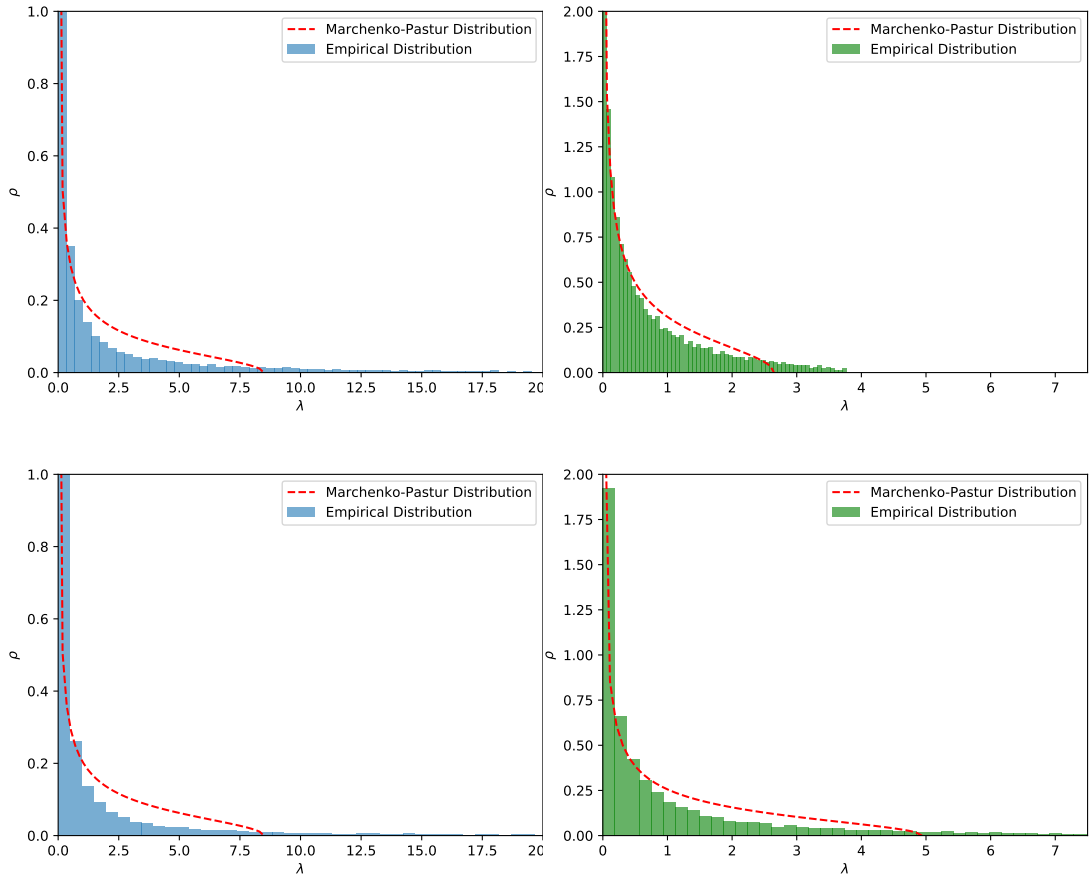


Figure 3.2: Examples of the distribution of the eigenvalues of $\mathbf{\Lambda}_{h^{(\ell)}}$ (left in blue) and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$ (right in green) at the 3rd (Above) and 7th layer (below). The dash red line corresponds to the Marchenko-Pastur distribution with the same mean as the spectrum.

The log-determinant of a matrix whose eigenvalues are distributed according to the Marchenko-Pastur distribution is defined as

$$\mathbb{E}_\mu[log(|X|)/n] = \int_{a_-}^{a_+} \frac{\log(x)}{2\pi\sigma^2} \frac{\sqrt{(\sigma^2 a_+ - x)(x - \sigma^2 a_-)}}{\gamma x} dx \tag{3.9}$$

$$= (1 - 1/\gamma)\log(1-\gamma) - 1 + \log(\sigma^2). \tag{3.10}$$

Consequently the log-determinant of a matrix, whose eigenvalues are distributed according to the Marchenko-Pastur distribution, is determined uniquely by the variance of the eigenvalues, σ^2, and the shape of the matrices γ. Since the vari-

ance of the eigenvalues is the same as their mean

$$\mathbb{E}_{\mu}[\lambda] = \frac{1}{2\pi\sigma^2}\int_0^{4\sigma^2}\sqrt{(4\sigma^2 - x)x}dx \tag{3.11}$$

$$= \frac{1}{2\pi\sigma^2}\int_{-\pi/2}^{\pi/2}\sqrt{(2\sigma^2 - 2\sigma^2\sin\theta)(2\sigma^2\sin\theta + 2\sigma^2)}2\sigma^2\cos\theta d\theta \tag{3.12}$$

$$= \frac{1}{2\pi\sigma^2}\int_{-\pi/2}^{\pi/2}4\sigma^4\sqrt{1-\sin^2\theta}\cos\theta d\theta = \frac{2\sigma^2}{\pi}\int_{-\pi/2}^{\pi/2}\cos^2\theta d\theta \tag{3.13}$$

$$= \frac{2\sigma^2}{\pi}\int_{-\pi/2}^{\pi/2}(1 + \frac{\cos(2\theta)}{2})d\theta = \sigma^2, \tag{3.14}$$

$q^{(\ell)}$ and $q_c^{(\ell)}$ are the expected eigenvalue of the matrices $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^{\top}\mathbf{\Sigma}_{xh^{(\ell)}}$ respectively, and $\gamma \to 1$, [1] it is possible to compute explicitly the log-determinant of $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^{\top}\mathbf{\Sigma}_{xh^{(\ell)}}$ under the assumption that the eigenvalues follow the Marchenko-Pastur distribution.

Proposition 3.1.2. *Assuming that the eigen-distributions of the matrices $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^{\top}\mathbf{\Sigma}_{xh^{(\ell)}}$ follow the Marchenko-Pastur distribution with means $q^{(\ell)}$ and $q_c^{(\ell)}$, the mutual information is lower bounded by*

$$I(\boldsymbol{x}; \boldsymbol{h}^{(\ell)}) \geq \frac{n}{2}\log\left(\frac{q^{(\ell)}}{q_c^{(\ell)}}\right) \tag{3.15}$$

Proof. From the previous chapter, the expected value of the eigenvalues of $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^{\top}\mathbf{\Sigma}_{xh^{(\ell)}}$ is known since the expected normalised trace is $\mathbb{E}_{\mathcal{W}:\ell}\left[\frac{tr\left(\mathbf{\Lambda}_{h^{(\ell)}}\right)}{n}\right] = q^{(\ell)}$ and $\mathbb{E}_{\mathcal{W}:\ell}\left[\frac{tr\left(\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^{\top}\mathbf{\Sigma}_{xh^{(\ell)}}\right)}{n}\right] = q_c^{(\ell)}$. Therefore, by also using equation (3.10) and $\gamma \to 1$, it follows that

$$I(\mathbf{x}; \mathbf{y}) \geq \frac{1}{2}\mathbb{E}_{\mathcal{W}:\ell}\left[\log\left(|\mathbf{\Lambda}_{h^{(\ell)}}|\right)\right] - \mathbb{E}_{\mathcal{W}:\ell}\left[\frac{1}{2}\log\left(\left|\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^{\top}\mathbf{\Sigma}_{xh^{(\ell)}}\right|\right)\right] \tag{3.16}$$

$$= \frac{n}{2}\log\left(q^{(\ell)}\right) - \frac{n}{2} - \frac{n}{2}\log\left(q_c^{(\ell)}\right) + \frac{n}{2} = \frac{n}{2}\log\left(\frac{q^{(\ell)}}{q_c^{(\ell)}}\right) \tag{3.17}$$

□

[1] If $\gamma = 1$ then the integral is not determined. This singularity is due to having no noise included in the distribution.

This shows that the lower bound defined under the mean-field approximation in Proposition 2.3.1 is equivalent to the one arising when considering the eigen-distribution of $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$ defined by the Marchenko-Pastur distribution.

3.1.2 The Stieltjes Transform

The approximation of the eigen-distributions of the various matrices with the Marchenko-Pastur distribution did not improve on the accuracy of the mutual information lower bound estimation. As a consequence, it is possible to rely on the Stieltjes transform to model accurately the eigen-spectrum of the matrices $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$. Previously, the Stieltjes transform was used to study the covariance matrix $\mathbf{\Lambda}_{h^{(2)}}$ of the first post-activation layer [6, 72]. In this chapter we are going to expand this kind of analysis to deeper layers of $\mathbf{\Lambda}_{h^{(\ell)}}$ and to any layer of $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$.

As was derived in [79], consider an $n \times n$ matrix $\mathbf{X}$, then its Stieltjes transform is given by its eigenvalues λ_k with the following expression

$$G_n^{\mathbf{X}}(z) = \frac{1}{n} tr\left((z\mathbf{I} - \mathbf{X})^{-1}\right) = \frac{1}{n}\sum_{k=1}^{n} \frac{1}{z - \lambda_k}, \tag{3.18}$$

where $\mathbf{I}$ is the identity matrix. When $\mathbf{X}$ is random, it is possible to define the empirical spectral distribution as

$$\rho_{\mathbf{X}}^n(\lambda) = \frac{1}{n}\sum_{k=1}^{n} \delta(\lambda - \lambda_k) \tag{3.19}$$

where $\delta(x)$ is the Dirac delta function. This allows to write the Stieltjes transform as

$$G_n^{\mathbf{X}}(z) = \int_{-\infty}^{+\infty} \frac{\rho_{\mathbf{X}}^n(\lambda)}{z - \lambda} d\lambda. \tag{3.20}$$

Since the Stieltjes transform is well behaved for $|z| \to \infty$, it is possible to rewrite the expression as

$$G_n^X(z) = \frac{1}{n}\sum_{k=1}^{n} \frac{1}{z - \lambda_k} = \frac{1}{zn}\sum_{k=1}^{n} \frac{1}{1 - \lambda_k/z} \tag{3.21}$$

$$= \frac{1}{n}\sum_{k=0}^{\infty}\frac{1}{z^{k+1}}\sum_{l=1}^{n}\lambda_l^k = \frac{1}{n}\sum_{k=0}^{\infty}\frac{1}{z^{k+1}}\frac{1}{n}tr(\mathbf{X}^k). \tag{3.22}$$

By considering the limit $n \to \infty$ the normalized traces of powers of $\mathbf{X}$ converge to their expectation values and therefore the Stieltjes transform converges to a deterministic limit $G^{\mathbf{X}}(z)$. By (3.2) this limiting Stieltjes transform is defined by the limiting eigen-distribution $\rho_{\mathbf{X}}$ as follows

$$G_{\mathbf{X}} = \int_{\mathbb{R}}\frac{\rho_{\mathbf{X}}(t)}{z-t}dt = \sum_{k=0}^{\infty}\frac{\mathfrak{m}_k}{z^{k+1}} \qquad z \in \mathbb{C}\setminus\mathbb{R} \tag{3.23}$$

where $\mathfrak{m}_k = \int \lambda^k \rho_{\mathbf{X}}(\lambda)d\lambda$. If the Stieltjes transform is known, it is also possible to retrieve the underlying distribution $\rho_{\mathbf{X}}$ by considering the Sokhotski-Plemelj formula

$$\rho_{\mathbf{X}}(t) = \frac{1}{\pi}\lim_{\nu\to 0^+} Im\left(G_{\mathbf{X}}(t-i\nu)\right). \tag{3.24}$$

The Stieltjes transform is especially useful when considering matrices that are free, a generalisation of independence in the matrix space. A matrix $\mathbf{M}$ is *free* if it has an eigen-spectrum whose distribution is rotationally invariant, meaning that the eigen-vectors are random orthogonal matrices; Wigner and Wishart matrices are two classic examples of free matrices. For more details on freeness, consider [59].

Following [79], to identify how free matrices interact the moment generating function $M_{\mathbf{X}}$ is introduced,

$$M_{\mathbf{X}}(z) = zG_{\mathbf{X}}(z) - 1 \tag{3.25}$$

and then the S-transform and the R-transform are defined as

$$S_{\mathbf{X}}(z) = \frac{1+z}{zM_{\mathbf{X}}^{-1}(z)}, \qquad R_{\mathbf{X}}(z) = G_{\mathbf{X}}^{-1}(z) - \frac{1}{z}. \tag{3.26}$$

Specifically, if two matrices $\mathbf{M}$ and $\mathbf{C}$ are free relative to each other, it is possible to compute the Stieltjes transform of $\mathbf{C}^{1/2}\mathbf{M}\mathbf{C}^{1/2}$, using the multiplicativity of the S-transform

$$S_{\mathbf{C}^\top\mathbf{M}\mathbf{C}}(z) = S_{\mathbf{C}^\top\mathbf{C}}(z)S_{\mathbf{M}}(z). \tag{3.27}$$

Furthermore, if $\mathbf{M}$ and $\mathbf{C}$ are free relative to each other, then it is possible to obtain the Stieltjes transform of $\mathbf{M}+\mathbf{C}$ via the additivity of the R-transform which is also based on the Stieltjes transform

$$R_{\mathbf{M}+\mathbf{C}}(z) = R_{\mathbf{M}}(z) + R_{\mathbf{C}}(z). \tag{3.28}$$

These properties of the Stieltjes transform of free matrices is going to be central to the modelling of the spectral distribution of $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^{\top}\mathbf{\Sigma}_{xh^{(\ell)}}$.

3.1.2.1 Computing the Stieltjes Transform of the Product of Free Matrices

As exposed in [79], from (3.26) and (3.27) it is possible to compute the inverse of the moment generating function of the the product $\mathbf{WAW}^{\top}$

$$M^{-1}_{\mathbf{WAW}}(z) = \frac{1+z}{zS_{\mathbf{WAW}}(z)} = \frac{1+z}{zS_{\mathbf{WW}}(z)S_{\mathbf{A}}(z)} \tag{3.29}$$

$$= \frac{1+z}{zS_{\mathbf{WW}}(z)}\frac{zM^{-1}_{\mathbf{A}}(z)}{1+z} = \frac{M^{-1}_{\mathbf{A}}(z)}{S_{\mathbf{WW}}(z)} \tag{3.30}$$

and generate an implicit relation between the moment generating functions $M_{\mathbf{WAW}}$ and $M_{\mathbf{A}}$. Since

$$M_{\mathbf{WAW}}(M^{-1}_{\mathbf{WAW}}(z)) = z = M_{\mathbf{A}}(M^{-1}_{\mathbf{A}}(z)) \tag{3.31}$$

$$= M_{\mathbf{A}}(S_{\mathbf{WW}}(z)M^{-1}_{\mathbf{WAW}}(z)) \tag{3.32}$$

by considering $\hat{z} = M^{-1}_{\mathbf{WAW}}(z)$, the following identity is retrieved

$$M_{\mathbf{WAW}}(\hat{z}) = M_{\mathbf{A}}(S_{\mathbf{WW}}(M_{\mathbf{WAW}}(\hat{z}))\hat{z}). \tag{3.33}$$

If $M_{\mathbf{A}}$ and $S_{\mathbf{WW}}$ are known, the algorithm to compute the density of the eigenspectrum of $\mathbf{WAW}$ is described in Algorithm 1; the property of the Stieltjes transform of behaving as $1/z$ for $|z| \to \infty$ is used.

Algorithm 1: Computing $\rho_{\mathbf{WAW}}(\lambda)$ [71]

Choose 2N steps and a dimension $b > 1$;
Initialize $z_0 = \lambda - ib^N$ and $G = 1/z$;
for $k \in \{1, ..., 2N\}$ **do**
  $z_k \leftarrow \lambda - ib^{N-k}$;
  $M_k \leftarrow$ Root of (3.33) nearest to $z_k G - 1$;
  $G \leftarrow (M_k + 1)/z_k$
end
Return $\frac{1}{\pi} Im\,(G)$

3.1.2.2 Computing the Stieltjes Transform of the Sum of Free Matrices

As also shown in [79], for the addition of two matrices $\mathbf{A}$ and $\mathbf{B}$, from the identity of the R-transform the following holds

$$G^{-1}_{\mathbf{A}}(z) + R_{\mathbf{B}}(z) = G^{-1}_{\mathbf{AB}}(z) \tag{3.34}$$

and thus

$$G_{\mathbf{AB}}\left(G^{-1}_{\mathbf{AB}}(z)\right) = z = G_{\mathbf{A}}\left(G^{-1}_{\mathbf{A}}(z)\right) \tag{3.35}$$

$$= G_{\mathbf{A}}\left(G^{-1}_{\mathbf{AB}}(z) - R_{\mathbf{B}}(z)\right). \tag{3.36}$$

By considering $\bar{z} = G^{-1}_{\mathbf{AB}}(z)$, the following implicit equation is obtained

$$G_{\mathbf{AB}}(\bar{z}) = G_{\mathbf{A}}\left(\bar{z} - R_{\mathbf{B}}(G_{\mathbf{AB}}(\bar{z}))\right) \tag{3.37}$$

and consequently the computation of $\rho_{\mathbf{A+B}}(x)$ is described in Algorithm 2.

Algorithm 2: Computing $\rho_{\mathbf{A+B}}(\lambda)$

Choose 2N steps and a dimension $b > 1$;
Initialize $z_0 = \lambda - ib^N$ and $G = 1/z$;
for $k \in \{1, ..., 2N\}$ **do**
  $z_k \leftarrow \lambda - ib^{N-k}$;
  $G \leftarrow$ Root of (3.37) nearest to G;
end
Return $\frac{1}{\pi} Im\,(G)$;

3.1.2.3 The Method of Moments

The method of moments was used in [72] and [6] to find an implicit recursive relation for the Stieltjes transform of matrix $\mathbf{\Lambda}_{h^{(2)}}$, $G_{\mathbf{\Lambda}_{h^{(2)}}}(z)$[2] in the limit of $n_\ell \to \infty$. Given a general random matrix $\mathbf{M}$, this method aims at evaluating the identity

$$G_{\mathbf{M}}(z) = \sum_{k=0}^{\infty} \frac{\mathfrak{m}_k}{z^{k+1}} \tag{3.38}$$

by determining exactly the moments, $\mathfrak{m}_k$, of the eigen-distribution $\rho_{\mathbf{M}}$ of $\mathbf{M}$. When considering the matrices relevant to the computation of the mutual information lower bound, the final expression of $\mathfrak{m}_k$ allows to identify the leading contribution and therefore to show that there is a unique measure yielding the identified transform; but the expression is too computationally complex to be computed in practice. However, by computing the moments $\mathfrak{m}_k$ it is possible to identify an implicit relation that the Stieltjes transform has to satisfy and that allows to compute the transform efficiently. Therefore, computing the moments will indirectly allow to identify the Stieltjes transform.

In this thesis similarly to [72] and [6], the method of moments is applied to matrices $\mathbf{M} \in \mathbb{R}^{n_1 \times n_1}$ that are the outer product of two matrices $\mathbf{M} = \frac{1}{m}\mathbf{Y}\mathbf{Y}^\top$ where the elements of $\mathbf{Y} \in \mathbb{R}^{n_1 \times m}$ that share an index are not independent, i.e. $\mathbf{Y}_{i,j} \perp\!\!\!\perp \mathbf{Y}_{p,q}$ but $\mathbf{Y}_{i,j} \not\perp\!\!\!\perp \mathbf{Y}_{i,q}$ and $\mathbf{Y}_{i,j} \not\perp\!\!\!\perp \mathbf{Y}_{p,j}$ for $i \neq p$ and $j \neq q$. Specifically, in this context the method of moments aims to identify $G(z)$ by considering the limiting moments $\mathfrak{m}_k = \lim_{n\to\infty} \mathfrak{m}_k^{(n)}$ [3] where

$$\mathfrak{m}_k^{(n)} = \frac{1}{n_1}\mathbb{E}[tr(\mathbf{M}^k)] = \frac{1}{n_1}\mathbb{E}\left[\sum_{i_1,\ldots,i_k \in [n]} \mathbf{M}_{i_1 i_2}\mathbf{M}_{i_2 i_3}\ldots\mathbf{M}_{i_k i_1}\right] \tag{3.39}$$

$$= \frac{1}{n_1 m^k}\mathbb{E}\left[\sum_{\substack{i_1,\ldots,i_k \in [n] \\ \mu_1,\ldots,\mu_k \in [m]}} \mathbf{Y}_{i_1\mu_1}\mathbf{Y}_{i_2\mu_1}\mathbf{Y}_{i_2\mu_2}\mathbf{Y}_{i_3\mu_2}\ldots\mathbf{Y}_{i_k\mu_k}\mathbf{Y}_{i_1\mu_k}\right] \tag{3.40}$$

[2] In the upcoming work the subscript on the Stieltjes transform identifying the corresponding matrix on which the transform is computed is going to be dropped unless the context requires it.

[3] Note that with the expression $n \to \infty$ we assume that all of the layer dimension of the nets are proportional to each other and therefore $n_\ell \to \infty$ for any ℓ.

$$= \frac{1}{n_1 m^k} \sum_{\substack{i_1,\ldots,i_k \in [n] \\ \mu_1,\ldots,\mu_k \in [m]}} \mathbb{E}\left[\mathbf{Y}_{i_1\mu_1}\mathbf{Y}_{i_2\mu_1}\mathbf{Y}_{i_2\mu_2}\mathbf{Y}_{i_3\mu_2}\ldots\mathbf{Y}_{i_k\mu_k}\mathbf{Y}_{i_1\mu_k}\right] \tag{3.41}$$

and these finite dimensional moments $\mathfrak{m}_k^{(n)}$ are computed by associating each sequence of indices $\{(i_1,\mu_1),\ldots,(i_1,\mu_k)\}$ to a pattern which is going to be represented with a graph. These graphs are such that despite having different indices, all the addends in (3.41) sharing the same pattern have the same expected value because of the independence of the weights. Therefore, the addends can be grouped according to their pattern, and, once the contribution of each pattern[4] is established, it is necessary to only quantify the pattern's frequency in the outer sum of (3.41).

The patterns that arise in (3.41) are identifiable as graphs by associating the indices, i_ξ and μ_ξ, to the vertex indices of a graph whose edges are defined by $Y_{i_\xi\mu_\xi}$. The work in [72] and [6] proves that (3.41) is dominated by the terms in the sum whose index pattern is associated to a connected outer-planar graphs in which all blocks are simple even cycles, and these graphs are defined as *admissible graphs*; for details consider Definition 3.1.1. For example, all the admissible patterns for $k = 3$ are shown in Figure 3.3; all the admissible graphs would consist in permutations of the indices.

Definition 3.1.1 ([72]). *For any positive integer k, a $2k$-cycle, is an admissible graph. Start by labelling the vertices in the $2k$-cycle as $1,\ldots,2k$ in a clockwise fashion. Consider any pair of vertices v_1 and v_2 of the same parity, one may obtain an admissible graph from another admissible graph by identifying v_1 and v_2 if there exist two vertex-disjoint paths between v_1 and v_2. The merged vertex is assigned the same parity as that on v_1 and v_2.*

The above definition of admissible graphs, allows to compute the moments $\mathfrak{m}_k^{(n)}$ for finite n-dimensional matrices $\mathbf{M}$ by first considering simple cycles and then aggregating their contributions. At leading order, the contribution of a graph is multiplicative in the contributions of its constituent cycles.

[4]From now on, the contribution of a pattern or a graph will correspond to the expected value of any term in (3.41) whose indices define the same pattern or graph.

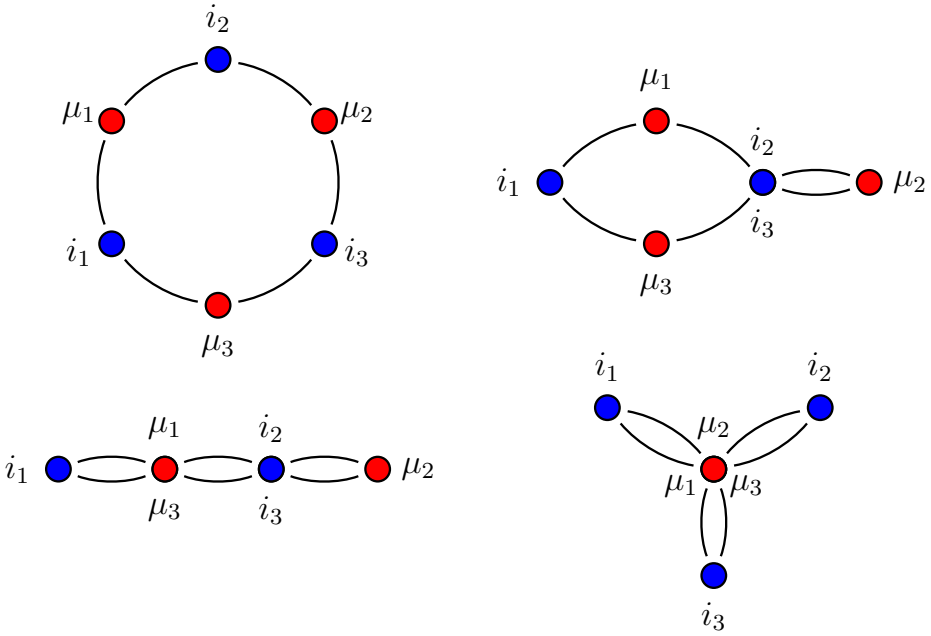


Figure 3.3: Admissible patterns according to Definition 3.1.1 when considering k=3. Any admissible graph with $k = 3$ consist of a permutation of the indices in the illustrated patterns. This Figures reproduces parts of [72, Figure S1].

When considering the limiting moment $\mathfrak{m}_k = \lim_{n\to\infty} \mathfrak{m}_k^{(n)}$, the explicit formulation of the moments allow to explicitly determine the Stieltjes transform of the matrices relevant to the mutual information lower bound estimation, however, this formulation is too complex to be used in practice. Nevertheless, thanks to the explicit formulation of the Stiletjes transform, it is possible to retrieve an implicit formulation of the Stieltjes transform that can be computed in practice. To find this implicit formulation, two properties of the limiting moments $\mathfrak{m}_k$ are considered: the fact that there exists a generating function based on $\mathfrak{m}_k$ that is related to the limiting Stieltjes transform G, and the fact that the multiplicativity of the constituent cycles used to compute $\mathfrak{m}_k^{(n)}$ is preserved.

3.1.3 Structure of Sections 3.2 and 3.3

The modelling of the matrices $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$ in Sections 3.2 and 3.3 is going to follow similar structures. Initially only the second layer is considered

in Sections 3.2.1 and 3.3.1; then the analyses are expanded to deeper layers in Section 3.2.2 and 3.3.2.

The study of the eigen-spectra of $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$ is going to be reduced to the understanding of respectively the matrices $\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell-1)})\phi(\mathbf{h}^{(\ell-1)})^\top]$ and $\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell-1)})\phi(\mathbf{h}^{(\ell-1)})^\top] - \mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell-1)})\mathbf{x}^\top]\mathbb{E}_{\mathbf{x}}[\mathbf{x}\phi(\mathbf{h}^{(\ell-1)})^\top]$ at the beginning of each relative section. The matrix $\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^1)\phi(\mathbf{h}^1)^\top]$ was studied in [72] when there is no bias and no noise, and therefore is included in Section 3.2.1. In Section 3.2.2 the extension to deeper layers in [6] is initially considered, but the inaccuracies lead to the development of a new model that performs more accurately.

Relative to $\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})\phi(\mathbf{h}^{(\ell)})^\top] - \mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})\mathbf{x}^\top]\mathbb{E}_{\mathbf{x}}[\mathbf{x}\phi(\mathbf{h}^{(\ell)})^\top]$ there was no prior work, therefore Section 3.3.1 focuses on establishing an understanding for the second layer. Then Subsection 3.3.2 expands these results to deeper layers; it will be shown that the technique used to obtain an exact approximation for $\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})\phi(\mathbf{h}^{(\ell)})^\top]$ are not amenable for this matrix, therefore some further assumption will be considered.

3.2 Eigenspectrum of $\mathbf{\Lambda}_{h^{(\ell)}}$

Given the weights $\mathbf{W}^{(\ell)}$ and $\mathbf{b}^{(\ell)}$, $\mathbf{\Lambda}_{h^{(\ell)}}$ is defined as

$$\Lambda_{\mathbf{h}^{(\ell+1)}} = \mathbb{E}_x[(\mathbf{W}^{(\ell)}\phi(\mathbf{h}^{(\ell)}) + \mathbf{n}^{(\ell)} + \mathbf{b}^{(\ell)})(\mathbf{W}^{(\ell)}\phi(\mathbf{h}^{(\ell)}) + \mathbf{n}^{(\ell)} + \mathbf{b}^{(\ell)})^\top] \tag{3.42}$$

$$= \left[\mathbf{W}^{(\ell)}, \mathbf{b}^{(\ell)} + \mathbf{n}^{(\ell)}\right] \begin{bmatrix} \mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})\phi(\mathbf{h}^{(\ell)})^\top] & \mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})] \\ \mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})]^\top & \mathbf{I} \end{bmatrix} \left[\mathbf{W}^{(\ell)}, \mathbf{b}^{(\ell)} + \mathbf{n}^{(\ell)}\right]^\top \tag{3.43}$$

Since, only odd activation functions ϕ are considered in this work, the vector $\mathbb{E}[\phi(\mathbf{h}^\ell)]$ is null and $\Lambda_{\mathbf{h}^{(\ell+1)}}$ can be rewritten as

$$\Lambda_{\mathbf{h}^{(\ell+1)}} = \frac{\sigma_w^2}{n}\boldsymbol{W}^{(\ell)}\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})\phi(\mathbf{h}^{(\ell)})^\top]\boldsymbol{W}^{(\ell)^\top} + (\mathbf{b}^{(\ell)} + \mathbf{n}^{(\ell)})(\mathbf{b}^{(\ell)} + \mathbf{n}^{(\ell)})^\top. \tag{3.44}$$

The matrix $\mathbf{W}^{(\ell)}\mathbf{W}^{(\ell)^\top}$ is free relative to $\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})\phi(\mathbf{h}^{(\ell)})^\top]$, therefore it is possible to study their Stieltjes transform thanks to the multiplicativity of the S-transform. It is possible to use Algorithm 1 thanks to the matrices $\mathbf{W}^{(\ell)}$ in (3.44) being Gaussian, for which $S_{\mathbf{WW}}(z) = \frac{1}{\sigma_w^2(1+z)}$, and thanks to the expression of the moment

generating function $M_{\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})\phi(\mathbf{h}^{(\ell)})^\top]}$ that can be computed numerically, as it will be shown in the next subsections. Moreover, since $\mathbf{W}^{(\ell)}\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})\phi(\mathbf{h}^{(\ell)})^\top]\mathbf{W}^{(\ell)^\top}$ is free relative to $(\mathbf{b}^{(\ell)}+\mathbf{n}^{(\ell)})(\mathbf{b}^{(\ell)}+\mathbf{n}^{(\ell)})^\top$, it is then possible to study the full spectrum of the matrix $\Lambda_{\mathbf{h}^{(\ell+1)}}$ thanks to the additivity of the R-transform and implement it thanks to Algorithm 2.

Therefore in this section we determine the Stieltjes transform of the matrix $\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})\phi(\mathbf{h}^{(\ell)})^\top]$ which is central to the computation of $\mathbf{\Lambda}_{h^{(\ell)}}$. In particular, first the results in [6, 72] relative to the second layer covariance are considered in Section 3.2.1. In the following Section 3.2.2 deeper layers are considered and after establishing that the Stieltjes transform identified in [6, 72] does not generalise well to deeper layers, a new approach is going to be proposed to model how signals correlate through the layers and thus correctly model the eigen-distribution of the covariance as deeper layers.

In the next sections, to consider the expectation over $\mathbf{x}$ and use the method of moments it is necessary to consider the expectation as a sample mean over $m \to \infty$ samples. Therefore, we will not consider an input vector $\mathbf{x}$ but an input matrix $\mathbf{X} = [\mathbf{x}_1, \ldots, \mathbf{x}_m]$ where the columns are m i.i.d. vectors $\mathbf{x}_i$.

3.2.1 Stieltjes Transform of the Matrix $\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(1)})\phi(\mathbf{h}^{(1)})^\top]$ with $\sigma_b = 0$ and $\sigma_n = 0$

The work in [72] addressed the problem of identifying the eigen-distribution of $\phi(\mathbf{XW})\phi(\mathbf{WX})^\top$ with a training capacity perspective[5]. For a two layers neural network, it is possible to identify the optimal weights of the second layer as a function of the Stieltjes transform of $\phi(\mathbf{WX})\phi(\mathbf{WX})^\top$ and then study the expected performance on the training set. The Stieltjes transform of $\phi(\mathbf{WX})\phi(\mathbf{WX})^\top$ is presented in Theorem 3.2.1 and when $m/n \to \infty$ then the Stieltjes transform of $\phi(\mathbf{WX})\phi(\mathbf{WX})^\top$ is the same as the one of $\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^1)\phi(\mathbf{h}^1)^\top]$.

Theorem 3.2.1 ([72, Theorem 1]). *Consider an odd activation function ϕ satisfying*

$$\left|\int \frac{dz}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \phi^k(\sigma_w \sigma_x z)\right| < \infty \quad \textit{and} \quad \left|\int \frac{dz}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \phi^{(k)}(\sigma_w \sigma_x z)\right| < \infty \tag{3.45}$$

[5] In this subsection the superscript $^{(\ell)}$ is omitted for clarity.

for $k > 1$ with $\phi^{(k)}$ being the k-th derivative of ϕ, an input $\boldsymbol{X} \in \mathbb{R}^{n_0 \times m}$ and $\boldsymbol{W} \in \mathbb{R}^{n_1 \times n_o}$ with their respective columns being sampled as follows $\boldsymbol{w}_i^{(\ell)} \sim \mathcal{N}(\boldsymbol{0}, \frac{\sigma_w^2}{n_{\ell-1}}\boldsymbol{I})$ and $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma_x^2 \boldsymbol{I})$ and with $\boldsymbol{Y} = \phi(\boldsymbol{W}\boldsymbol{X})$ and

$$\psi = \frac{n_0}{m}, \quad \varphi = \frac{n_0}{n_1}. \tag{3.46}$$

Then the eigen-distribution of the empirical covariance matrix $\boldsymbol{M} = \frac{1}{m}\boldsymbol{Y}\boldsymbol{Y}^\top$ is asymptotically defined by

$$G_{\boldsymbol{M}}(z) \simeq \frac{\varphi}{z} H(\varphi z) + \frac{1-\varphi}{z} \tag{3.47}$$

where the generating function $H(z)$ satisfies the following recursive relation

$$H(z) = 1 + \frac{(\theta_1 - \theta_2) H_\varphi(z) H_\psi(z)}{z} + \frac{H_\varphi(z) H_\psi(z) \theta_2}{z - H_\varphi(z) H_\psi(z) \theta_2} \tag{3.48}$$

and

$$H_\psi(z) = 1 + (H(z) - 1)\psi, \quad H_\varphi(z) = 1 + (H(z) - 1)\varphi \tag{3.49}$$

and

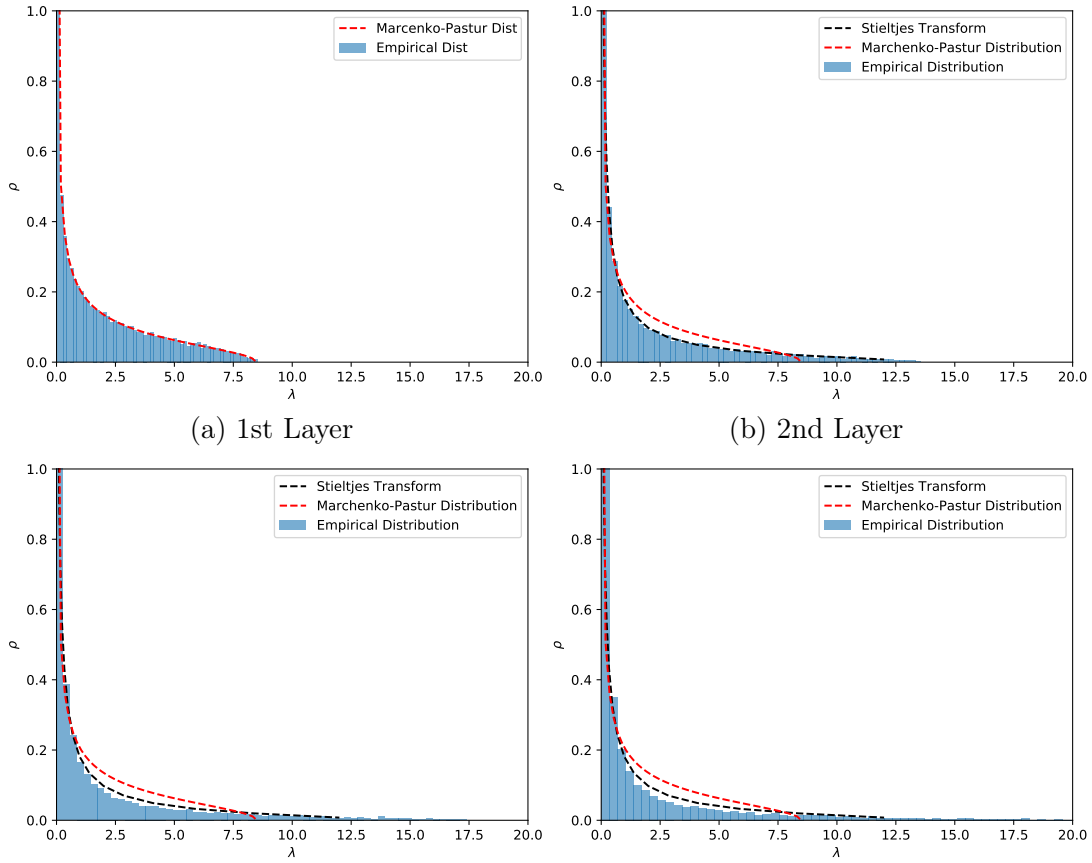
$$\theta_1 = \int_{\mathbb{R}} \phi(\sigma_w \sigma_x t)^2 \mathcal{D}t \qquad \textit{and} \qquad \theta_2 = \left(\sigma_w \sigma_x \int_{\mathbb{R}} \phi'(\sigma_w \sigma_x t) \mathcal{D}t \right)^2. \tag{3.50}$$

In the mutual information problem $\psi = 0$ and $\varphi = 1$, and this implies that the moment generating function of the covariance matrix of the first post-activation hidden layer satisfies $M_{\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^1)\phi(\mathbf{h}^1)^\top]}(z) = H(z) - 1$ with $H(z)$ being such that

$$H(z) = 1 + \frac{(\theta_1 - \theta_2) H(z)}{z} + \frac{H(z)\theta_2}{z - H(z)\theta_2}. \tag{3.51}$$

3.2.2 Stieltjes Transform of the Matrix $\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})\phi(\mathbf{h}^{(\ell)})^\top]$ with $\sigma_b = 0$ and $\sigma_n = 0$

The work in [6], showed that the same expression of the Stieltjes transform in Theorem 3.2.1 holds also for the case in which the input has sub-exponential distribution in the large limit n. Since the activation function, ϕ, is upper and lower bounded, the distribution of the post-activation hidden layers is sub-Gaussian and



(a) 1st Layer (b) 2nd Layer

(c) 3rd Layer (d) 4th Layer

Figure 3.4: Approximation of the empirical eigen-spectrum for the matrix $\mathbf{\Lambda}_{h^{(\ell)}}$ with $n = 100$ and $(\sigma_w, \sigma_b, \sigma_n) = (2, 0, 0.001)$ at the first four layers. The Stieltjes transform induced approximation models the tail of the distribution more accurately than the Marchenko-Pastur approximation does, especially for the second layer where the approximation seems to fit the empirical distribution exactly.

the observation in [6] allows to compute the spectral distribution of the covariance matrices after the first layer. Specifically, by considering

$$\mathbf{Y}^{(\ell+1)} = \phi\left(\mathbf{W}\mathbf{Y}^{(\ell)}\right), \qquad \mathbf{M}^{(\ell+1)} = \frac{1}{m}\mathbf{Y}^{(\ell+1)}\mathbf{Y}^{(\ell+1)\top}, \tag{3.52}$$

[6] suggests to implement Theorem 3.2.1 with $\mathbf{X} = \mathbf{Y}^{(\ell)}$ and $\sigma_x^2 = q^{(\ell)}/\sigma_w^2$.

In Figure 3.4 the distribution generated by considering the Stieltes transform of the matrices $\mathbf{W}^{(\ell)}\mathbf{Y}^{(\ell)}$ is compared to the empirical results when $\phi(\cdot) = tanh(\cdot)$ and $(\sigma_w, \sigma_b, \sigma_n) = (2, 0, 0.01)$. It is possible to observe that for the second layer,

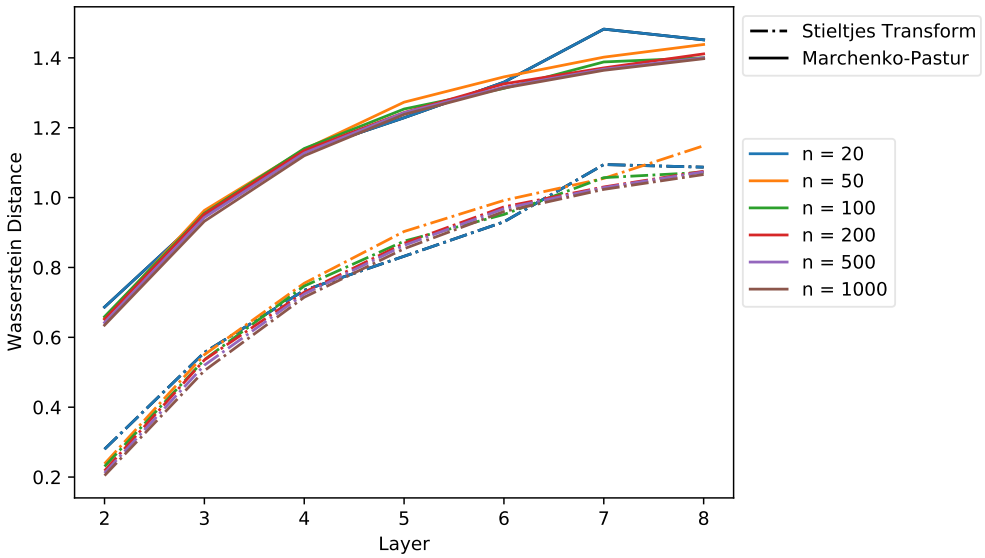


Figure 3.5: Comparison of the Wasserstein distance between the empirical distribution of $\mathbf{\Lambda}_{h^{(\ell)}}$ for different network dimensions and the distribution attained with the Stieltjes transform (dash-dotted) and with the Marchenko-Pastur distribution (solid). For the second layer the Wasserstein distance decreases when n increases as expected.

the distribution retrieved from the implicit formulation of the Stieltjes transform in [72] is accurately mapping the empirical distribution. However, as the layers increase there is an error in the approximation. To understand if the error is due to the dimensionality of the problem n, since [6] prove that the sub-gaussianity is sufficient in the large n limit, in Figure 3.5 the Wasserstein distance between the empirical distributions and the ones obtained with the Stieltjes transform is plotted in the setting with $(\sigma_w, \sigma_b, \sigma_n) = (2, 0, 0.01)$. The results show that there is an error in the approximation which gets worse as the layers increase.

As it was suggested in [75], when propagating through a network, signals might become correlated with each other. Therefore, the deeper the layers, the more correlated the entries in the hidden vectors might become. It will be shown with the results in this subsection that the lack of modelling of the correlation in Equations (3.52) is the primary reason of the discrepancy with the empirical results illustrated in Figures 3.4 and 3.5. We retrieve a recursive equation for the Stieltjes transform when the input has correlated elements. To do this, we consider the case where

the post-activation layer samples are defined by

$$\mathbf{Y} = \phi(\mathbf{W}\tilde{\mathbf{X}}) \tag{3.53}$$

with $\tilde{\mathbf{X}} \in \mathbb{R}^{n_0 \times m}$, $\mathbf{W} \in \mathbb{R}^{n_1 \times n_0}$, and $\mathbf{M} = \frac{1}{m}\mathbf{Y}\mathbf{Y}^\top$. Specifically, the column vectors $\tilde{\mathbf{x}}_i$ are considered to be identically independently distributed according to a Gaussian distribution $\tilde{\mathbf{x}}_i \sim \mathcal{N}(\mathbf{0}, \sigma_x^2 \mathbf{O}\mathbf{\Sigma}\mathbf{O}^\top)$. By considering a matrix $\mathbf{X}$ whose columns are $\mathbf{x}_i \sim \mathcal{N}(\mathbf{0}, \sigma_x^2\mathbf{I})$, since $\mathbf{W}$ is also a Gaussian matrix for which $\mathbf{W} \sim \mathbf{W}\mathbf{O}$ and $\tilde{\mathbf{x}}_i \sim \mathbf{O}\mathbf{\Sigma}^{1/2}\mathbf{x}_i$, in the upcoming analysis $\mathbf{W}\mathbf{\Sigma}^{1/2}\mathbf{X}$ will be considered rather than $\mathbf{W}\tilde{\mathbf{X}}$. Note that if the post-activation layer is assumed to be normally distributed, then this same analysis could be applied at deeper layers by considering $\mathbf{X} = \mathbf{Y}^{(\ell)}$ and considering $\sigma_x^2 = q^{(\ell)}/\sigma_w^2$.

In this setting we find an implicit equation for the Stieltjes transform of $\mathbf{M}$ that is defined in Theorem 3.2.2.

Theorem 3.2.2. *Consider the odd activation function ϕ*

$$\left| \int \frac{dz}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \phi^k(\sigma_w \sigma_x z) \right| < \infty \quad \textit{and} \quad \left| \int \frac{dz}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \phi^{(k)}(\sigma_w \sigma_x z) \right| < \infty \tag{3.54}$$

for $k > 1$ with $\phi^{(k)}$ being the k-th derivative of ϕ and the matrices $\boldsymbol{W} \in \mathbb{R}^{n_1 \times n_0}$ and $\tilde{\boldsymbol{X}} \in \mathbb{R}^{n_0 \times m}$ with their respective columns being sampled as follows $\boldsymbol{w}_i \sim \mathcal{N}(\boldsymbol{0}, \frac{\sigma_w^2}{n}\boldsymbol{I})$ and $\tilde{\boldsymbol{x}}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma_x^2 \boldsymbol{O}\mathbf{\Sigma}\boldsymbol{O}^\top)$ with the eigen-spectrum of $\mathbf{\Sigma}$ being bounded, with $\tilde{\mu}_k = \int \lambda^k \rho_{\mathbf{\Sigma}}(\lambda) d\lambda$. Then, defining $\psi = \frac{n_0}{m}$ and $\varphi = \frac{n_0}{n_1}$, the matrix $\boldsymbol{M} = \frac{1}{m}\boldsymbol{Y}\boldsymbol{Y}^\top$ with $\boldsymbol{Y} = \phi(\boldsymbol{W}\tilde{\boldsymbol{X}})$ has an eigen-distribution whose Stieltjes transform is asymptotically defined by

$$G_{\boldsymbol{M}}(z) \simeq \frac{1-\varphi}{z} + \frac{\varphi}{z} H(z) \tag{3.55}$$

where

$$H(z) = \frac{\tilde{\theta}_1 - \tilde{\mu}_1 \tilde{\theta}_2}{z} H_\psi(z) H_\varphi(z) + \frac{z}{\tilde{\theta}_2 H_\psi(z) H_\varphi(z)} G_{\mathbf{\Sigma}}\left(\frac{z}{\tilde{\theta}_2 H_\psi(z) H_\varphi(z)} \right) \tag{3.56}$$

with

$$\tilde{\theta}_1 = \int \frac{1}{\sqrt{2\pi}} \phi(\sqrt{\tilde{\mu}_1}\sigma_w \sigma_x \tilde{z})^2 e^{-\frac{\tilde{z}^2}{2}} d\tilde{z}, \tag{3.57}$$

$$\tilde{\theta}_2 = \left(\int \frac{\sigma_w \sigma_x}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \phi'(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1}\tilde{z}) d\tilde{z} \right)^2, \tag{3.58}$$

$H_\psi(z) = 1 + \psi(H(z) - 1)$, $H_\varphi = 1 + \varphi(H(z) - 1)$, and $G_{\mathbf{\Sigma}}(z) = \sum_{k=0}^{\infty} \frac{\tilde{\mu}_k}{z^{k+1}}$.

This theorem is going to be proven within the next subsections following the structure suggested at the end of Section 3.1.2.3. In Subsection 3.2.2.1 the contribution of single cycle patterns are determined and in Subsection 3.2.2.2 they are aggregated to identify the contribution to the moments of an admissible graph. This allows to define a generating function and to determine the contribution of two graphs that are connected via a vertex. These observations are then finally used in Subsection 3.2.2.3 to prove Theorem 3.2.2 and show its consistency with Theorem 3.2.1 if the inputs are considered to be independent, i.e. $\mathbf{\Sigma} = \mathbf{I}$.

3.2.2.1 $2k$-cycle Contributions

Consider cycles of length $2k$, i.e. in Equation (3.41) $i_1 \neq i_2 \neq ... \neq i_k$ and $\mu_1 \neq \mu_2 \neq ... \neq \mu_k$; then the expected contribution is

$$E_{2k}^{(n)} = \mathbb{E}\left[\mathbf{Y}_{i_1\mu_1}\mathbf{Y}_{i_2\mu_1}...\mathbf{Y}_{\mu_k i_1}\right] \tag{3.59}$$

$$= \int \phi(\sum_l \mathbf{W}_{i_1 l}\mathbf{\Sigma}_{ll}^{1/2}\mathbf{X}_{l\mu 1})\phi(\sum_l \mathbf{W}_{i_2 l}\mathbf{\Sigma}_{ll}^{1/2}\mathbf{X}_{l\mu 1})...\phi(\sum_l \mathbf{W}_{i_1 l}\mathbf{\Sigma}_{ll}^{1/2}\mathbf{X}_{l\mu_k})\mathcal{D}\mathbf{W}\mathcal{D}\mathbf{X}\mathcal{D}\mathbf{\Sigma} \tag{3.60}$$

where

$$\mathcal{D}\mathbf{W} = \prod_{i,j=1}^{n} \frac{d\mathbf{W}_{ij}}{\sqrt{2\pi\sigma_w^2/n}} e^{-\frac{n\mathbf{W}_{ij}^2}{2\sigma_w^2}} \quad \text{and} \quad \mathcal{D}\mathbf{X} = \prod_{i,\mu=1}^{n} \frac{d\mathbf{X}_{i\mu}}{\sqrt{2\pi\sigma_x^2}} e^{-\frac{\mathbf{X}_{i\mu}^2}{2\sigma_x^2}} \tag{3.61}$$

and $\mathcal{D}\mathbf{\Sigma}$ is the distribution of the eigenvalues of the covariance matrix.

It is necessary to distinguish two cases, when $k = 1$ and when $k > 1$. In the former case

Lemma 3.2.3. *For $\boldsymbol{M} = \frac{1}{m}\boldsymbol{Y}\boldsymbol{Y}^\top$ and $\boldsymbol{Y} = \phi\left(\boldsymbol{W}\tilde{\boldsymbol{X}}\right)$ as defined in Theorem 3.2.2, then for $k = 1$*

$$E_2^{(n)} = \int \tilde{\theta}_1^{(n)} \mathcal{D}\mathbf{\Sigma}\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \tag{3.62}$$

with

$$\tilde{\theta}_1^{(n)} = \int \frac{1}{\sqrt{2\pi}} \phi(\sqrt{\tilde{\mu}_1^{(n)}}\sigma_w\sigma_x\tilde{z})^2 e^{-\frac{\tilde{z}^2}{2}} d\tilde{z} \tag{3.63}$$

where $\tilde{\mu}_k^{(n)} = tr(\mathbf{\Sigma}^k)/n$.

Proof. See proof in Section A.1.2. □

While in the latter case, when $k > 1$, the contribution $E_{2k}^{(n)}$ is computed with the following Lemma:

Lemma 3.2.4. *For* $\boldsymbol{M} = \frac{1}{m}\boldsymbol{Y}\boldsymbol{Y}^\top$ *and* $\boldsymbol{Y} = \phi\left(\boldsymbol{W}\tilde{\boldsymbol{X}}\right)$ *as defined in Theorem 3.2.2, then for* $k > 1$

$$E_{2k}^{(n)} = n^{1-k}\int \tilde{\theta_2}^k \tilde{\mu}_k^{(n)} \mathcal{D}\boldsymbol{\Sigma}\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \tag{3.64}$$

with

$$\tilde{\theta}_2^{(n)} = \left(\int \frac{\sigma_w\sigma_x}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \phi'(\sigma_w\sigma_x\sqrt{\tilde{\mu}_1^{(n)}}\tilde{z})d\tilde{z}\right)^2 \tag{3.65}$$

where $\tilde{\mu}_k^{(n)} = tr(\boldsymbol{\Sigma}^k)/n$.

Proof. See proof in Section A.1.3 □

3.2.2.2 Agglomeration of the Contributions

To compute the moments of the distribution, it is necessary to identify all of the possible patterns, identify their cardinality (each pattern can be achieved by selecting different vertices), and their expected contribution thanks to the previous subsection.

To count the different patterns, after naming I_i and I_μ the identifications in Definition 3.1.1 of respectively odd and even parity, the following variable is considered similarly to [6, 72]:

Definition 3.2.1. $\mathcal{B}(\overrightarrow{k}, I_i, I_\mu, b)$ *is the number of admissable pattern with* $2k$ *edges,* I_i *i-identifications,* I_μ *μ-identifications, c cycles of dimension* $\overrightarrow{k} = \{\tilde{k}_1, \ldots, \tilde{k}_c\}$, *and with exactly b cycles of size 2.*

Thanks to the formulation for the contribution of $2k$-cycles, the k-th moment of the eigen-distribution is determined in Proposition 3.2.5.

Proposition 3.2.5. *For* $\boldsymbol{M} = \frac{1}{m}\boldsymbol{Y}\boldsymbol{Y}^\top$ *and* $\boldsymbol{Y} = \phi\left(\boldsymbol{W}\tilde{\boldsymbol{X}}\right)$ *as defined in Theorem 3.2.2, the k-th moment of the eigen-distribution of* $\boldsymbol{M}$*, for* $k > 1$*, is*

$$\mathfrak{m}_k^{(n)} = \mathfrak{m}_k\left(1 + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right) \tag{3.66}$$

with

$$\mathfrak{m}_k = \lim_{n\to\infty} \mathfrak{m}_k^{(n)} \tag{3.67}$$

$$= \varphi^{1-k} \sum_{I_i,I_\mu=1}^{k} \sum_{b=0}^{I_i+I_\mu+1} \frac{\mathcal{A}(k, I_i, I_\mu, b)\tilde{\theta}_1^b \tilde{\theta}_2^{k-b} \varphi^{I_1}\psi^{I_\mu}}{2\pi((k-I_i)(k-I_\mu))^{\frac{1}{2}}} \left(\frac{e}{k-I_i}\right)^{k-I_1} \left(\frac{e}{k-I_\mu}\right)^{k-I_\mu} \tag{3.68}$$

$$= \varphi^{1-k} H_k. \tag{3.69}$$

where

$$\mathcal{A}(k, I_i, I_\mu, b) = \sum_{\overrightarrow{k}} \mathcal{B}(\overrightarrow{k}, I_i, I_\mu, b)\left(\prod_{c=1}^{1+I_i+I_\mu-b} \tilde{\mu}_{\tilde{k}_c}\right), \tag{3.70}$$

and

$$\tilde{\theta}_1 = \lim_{n\to\infty} \int \tilde{\theta}_1^{(n)} \mathcal{D}\boldsymbol{\Sigma} \tag{3.71}$$

$$\tilde{\theta}_2 = \lim_{n\to\infty} \int \tilde{\theta}_2^{(n)} \mathcal{D}\boldsymbol{\Sigma}, \tag{3.72}$$

and the new variable

$$H_k = \sum_{I_i,I_\mu=1}^{k} \sum_{b=0}^{I_i+I_\mu+1} \frac{\mathcal{A}(k, I_i, I_\mu, b)\tilde{\theta}_1^b \tilde{\theta}_2^{k-b} \varphi^{I_1}\psi^{I_\mu}}{2\pi((k-I_i)(k-I_\mu))^{\frac{1}{2}}} \left(\frac{e}{k-I_i}\right)^{k-I_1} \left(\frac{e}{k-I_\mu}\right)^{k-I_\mu}$$

is introduced.

Proof. See proof in Section A.1.4. □

This allows to identify the Stieltjes transform by introducing the generating function H.

Corollary 3.2.5.1. *For* $\boldsymbol{M} = \frac{1}{m}\boldsymbol{Y}\boldsymbol{Y}^\top$ *and* $\boldsymbol{Y} = \phi\left(\boldsymbol{W}\tilde{\boldsymbol{X}}\right)$ *as defined in Theorem 3.2.2, the Stieltjes transform of the eigen-distribution of* $\boldsymbol{M}$ *asymptotically satisfies the following equation*

$$G_{\boldsymbol{M}}(z) \simeq \frac{1-\varphi}{z} + \frac{\varphi}{z}H(z) \tag{3.73}$$

where $H(z)$ *is the generating function defined as* $H(z) = 1 + \sum_{k=1}^{\infty}\frac{H_k}{z^k\varphi^k} = 1 + \frac{1}{\varphi}\sum_{k=1}^{\infty}\frac{\mathfrak{m}_k}{z^k}$.

Proof. To start with we notice that using the definition of the Stieltjes transform (3.38)

$$G_{\mathbf{M}}(z) = \sum_{k=0}^{\infty}\frac{\mathfrak{m}_k^{(n)}}{z^{k+1}} = \sum_{k=0}^{\infty}\frac{\mathfrak{m}_k}{z^{k+1}}\left(1 + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right) \tag{3.74}$$

$$= G(z)\left(1 + \mathcal{O}\left(\frac{1}{\sqrt{n}}\right)\right) \tag{3.75}$$

By relying on equations (3.38) and (3.69)

$$G(z) = \frac{1}{z} + \sum_{k=1}^{\infty}\frac{\mathfrak{m}_k}{z^{k+1}} = \frac{1}{z} + \frac{\varphi}{z}\sum_{k=1}^{\infty}\frac{H_k}{z^k\varphi^k} \tag{3.76}$$

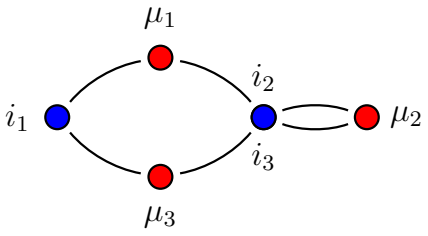
$$= \frac{1-\varphi}{z} + \frac{\varphi}{z}\left(1 + \sum_{k=1}^{\infty}\frac{H_k}{z^k\varphi^k}\right) = \frac{1-\varphi}{z} + \frac{\varphi}{z}H(z) \tag{3.77}$$

□

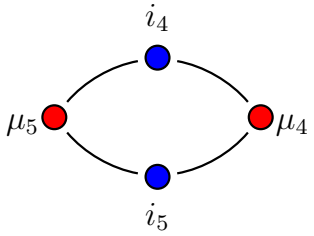
By defining the total moment $\mathfrak{m}_k$ in function of the contribution of all the graphs G that have $2k$ edges, i.e. $\mathfrak{m}_k = \sum_{G s.t. |G|=2k}\mathfrak{m}_k^G$, Proposition 3.2.5 also allows to formalise how the contribution of a graph is the product of the marginal contributions of the constituent sub-graphs with a correction term.

Corollary 3.2.5.2. *Consider* $\boldsymbol{M} = \frac{1}{m}\boldsymbol{Y}\boldsymbol{Y}^\top$ *and* $\boldsymbol{Y} = \phi\left(\boldsymbol{W}\tilde{\boldsymbol{X}}\right)$ *as defined in Theorem 3.2.2, and consider a graph* G_3 *with* $2(p+q)$ *edges that is defined, via either an* i *or* μ *identifications, by two graphs* G_1 *and* G_2 *with* $2p$ *and* $2q$ *edges. Then the asymptotic contribution,* $\mathfrak{m}_{p+q}^{G_3}$*, of the graph* G_3 *to the moment* $\mathfrak{m}_{p+q}$ *is equal to the product of the contributions of the block cycle graphs that define* G_3 *with a correction term. This is, if they are joined with an* i*-identification*

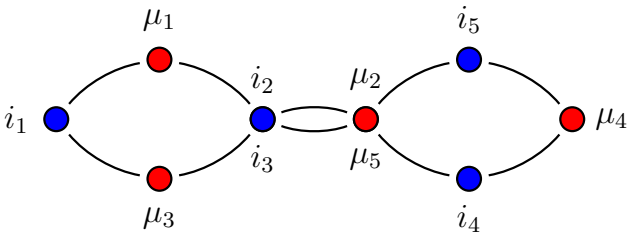
$$\mathfrak{m}_{q+p}^{G_3} = \mathfrak{m}_q^{G_1}\cdot\frac{n_1}{n_1}\mathfrak{m}_p^{G_2} \tag{3.78}$$



$$\mathfrak{m}_3^{G_1} = \frac{1}{n_1 m^3} \sum_{i_1=1}^{n_1} \tilde{\theta}_1 \tilde{\theta}_2^2 \tilde{\mu}_2 m^3 n_1 n_0^{-1} = \frac{1}{\varphi} \tilde{\theta}_1 \tilde{\theta}_2^2 \tilde{\mu}_2$$



$$\mathfrak{m}_2^{G_2} = \frac{1}{n_1 m^2} \sum_{i_1=1}^{n_1} \tilde{\theta}_2^2 \tilde{\mu}_2 m^2 n_1 n_0^{-1} = \frac{1}{\varphi} \tilde{\theta}_2^2 \tilde{\mu}_2$$



$$\mathfrak{m}_5^{G_3} = \frac{1}{n_1 m^5} \sum_{i_1=1}^{n_1} \tilde{\theta}_1 \tilde{\theta}_2^4 \tilde{\mu}_2^2 m^4 n_1^3 n_0^{-2} = \tilde{\theta}_1 \tilde{\theta}_2^4 \tilde{\mu}_2^2 m^{-1} n_1^3 n_0^{-2} = \frac{1}{\varphi^3} \phi \tilde{\theta}_1 \tilde{\theta}_2^4 \tilde{\mu}_2^2$$
$$= \frac{1}{\varphi} \tilde{\theta}_1 \tilde{\theta}_2^2 \tilde{\mu}_2 \cdot \frac{\phi}{\varphi^2} \tilde{\theta}_2^2 \tilde{\mu}_2 = \mathfrak{m}_3^{G_1} \cdot \frac{\phi}{\varphi} \mathfrak{m}_2^{G_2} = \mathfrak{m}_3^{G_1} \cdot \frac{n_1}{m} \mathfrak{m}_2^{G_2}$$

Figure 3.6: Example on how the contribution of two graphs G_1 and G_2 determines the contribution of their combination G_3 with a μ identification. The expressions of the marginal contributions is related to Proposition A.1.4. Note, with an I_i identification, we would have obtained that $\mathfrak{m}_5^{G_3} = \mathfrak{m}_3^{G_1} n_1 / n_1 \mathfrak{m}_2^{G_2}$.

and if they are joined with a μ-identification

$$\mathfrak{m}_{q+p}^{G_3} = \mathfrak{m}_q^{G_1} \frac{n_1}{m} \mathfrak{m}_p^{G_2}. \tag{3.79}$$

Consider Figure 3.6 for a visual interpretation of Corollary 3.2.5.2. The correction term in the denominator, n_1 and m, is a result of a further identification which entails a factor of either n_1 or m combinations less. While the correction term in the nominator is due to the extra $\frac{1}{n_1}$ in the moment expression (3.41).

3.2.2.3 Proof of the Implicit Relation in Theorem 3.2.2

The generating function $H(1/z) = \sum_{k=0}^{\infty} \frac{H_k}{\varphi^k} z^k$ corresponds besides for a constant term to the moment generating function of the eigen-distribution, which is uniquely defined as shown in Appendix A.1.5; it is a power series where the coefficients define the kth moment of the eigenvalue distribution according to

$$H_k/\varphi^k = \mathfrak{m}_k/\varphi.$$

The following proof will define an implicit formula for $H(z)$ that can be computed via root-finding methods.

Proof. The different moments of the eigen-distribution are computed as the aggregation of the contributions of different admissible graphs. Specifically, it was shown in Corollary 3.2.5.2 that given a graph G_3 defined by an i- or μ-identification of two admissible graphs G_1 and G_2, then the total contribution to the moment is

$$\mathfrak{m}_{q+p}^{G_3} = \mathfrak{m}_q^{G_1} \frac{n_1}{n_1} \mathfrak{m}_p^{G_2} \tag{3.80}$$

if the graphs are joined with an i-identification, while it is

$$\mathfrak{m}_{q+p}^{G_3} = \mathfrak{m}_q^{G_1} \frac{n_1}{m} \mathfrak{m}_p^{G_2}. \tag{3.81}$$

if they are joined with a μ-identification.

If a vertex v is fixed, we can use the moment generating function to define the contributions of all the possible patterns which include the chosen vertex. For each of these patterns, the vertex is going to be in a cycle of dimension 2ℓ, with a contribution $\mathfrak{m}_\ell^{2\ell-cycle}$, and to each of the remaining vertices either nothing or a further graph is connected. By considering the limiting valriables $E_{2\ell} = \lim_{n\to\infty} E_{2\ell}^{(n)}$, the total contribution, $C_{2\ell}$, to the different moments for graphs in which the vertex v is included in a 2ℓ cycle is defined by

$$C_{2\ell} = \frac{\mathfrak{m}_\ell^{2\ell-cycle}}{z^\ell} \left(1 + \sum_{G_\alpha} \frac{\mathfrak{m}_{|G_\alpha|}^{G_\alpha}}{z^{|G_\alpha|}}\right)^\ell \left(1 + \sum_{G_\alpha} \frac{n_1}{m} \frac{\mathfrak{m}_{|G_\alpha|}^{G_\alpha}}{z^{|G_\alpha|}}\right)^\ell \tag{3.82}$$

$$= \frac{1}{n_1 m^\ell z^\ell} E_{2\ell} n_1^{\ell-1} m^\ell \left(1 + \sum_{G_\alpha} \frac{\mathfrak{m}_{|G_\alpha|}^{G_\alpha}}{z^{|G_\alpha|}}\right)^\ell \left(1 + \sum_{G_\alpha} \frac{n_1}{m} \frac{\mathfrak{m}_{|G_\alpha|}^{G_\alpha}}{z^{|G_\alpha|}}\right)^\ell \tag{3.83}$$

$$= \frac{1}{n_1 m^\ell z^\ell} E_{2\ell} n_1^{\ell-1} m^\ell \left(1 + \frac{n_1}{n_1}\varphi(H(z)-1)\right)^\ell \left(1 + \frac{n_1}{m}\varphi(H(z)-1)\right)^\ell \quad (3.84)$$

$$= \frac{1}{n_1 m^\ell z^\ell} E_{2\ell} n_1^{\ell-1} m^\ell \left(1 + \varphi(H(z)-1)\right)^\ell \left(1 + \psi(H(z)-1)\right)^\ell \quad (3.85)$$

$$= \frac{1}{n_1 z^\ell} E_{2\ell} n_1^{\ell-1} H_\varphi^\ell H_\psi^\ell \quad (3.86)$$

where G_α is any possible graph, $|G_\alpha|$ is its dimension, we used the definition of the generating function

$$H(z) - 1 = \frac{1}{\varphi} \sum_{G_\alpha} \frac{\mathfrak{m}_{|G_\alpha|}^{G_\alpha}}{z^{|G_\alpha|}} \quad (3.87)$$

which is a reformulation of the definition in Corollary 3.2.5.1 by considering that $\mathfrak{m}_k = \sum_{Gs.t.|G|=2k} \mathfrak{m}_k^G$, and we defined

$$H_\varphi(z) = 1 + \varphi(H(z)-1) \quad \text{and} \quad H_\psi(z) = 1 + \psi(H(z)-1). \quad (3.88)$$

See Figure 3.7 for a visual interpretation.

Therefore, by summing the contribution $C_{2\ell}$ over all the possible fixed vertices v and over all the possible cycle dimensions ℓ, we retrieve the moment generating function.

$$\varphi(H(z)-1) = \sum_k \frac{\mathfrak{m}_k}{z^k} = \sum_{i=1}^{n_1} \sum_{\ell=1}^{\infty} C_{2\ell} \quad (3.89)$$

$$= \sum_{\ell=1}^{\infty} \frac{E_{2\ell} n_1^{\ell-1}}{z^\ell} H_\psi(z)^\ell H_\varphi(z)^\ell \quad (3.90)$$

$$= \frac{\tilde{\theta}_1}{z} H_\psi(z) H_\varphi(z) + \sum_{\ell=2}^{\infty} n_0^{1-\ell} \frac{\tilde{\theta}_2^\ell \tilde{\mu}_\ell}{z^\ell} n_1^{\ell-1} H_\psi(z)^\ell H_\varphi(z)^\ell \quad (3.91)$$

$$= \frac{\tilde{\theta}_1}{z} H_\psi(z) H_\varphi(z) + \sum_{\ell=2}^{\infty} \frac{n_0^{1-\ell}}{n_1^{1-\ell}} \frac{\tilde{\theta}_2^\ell \tilde{\mu}_\ell}{z^\ell} H_\psi(z)^\ell H_\varphi(z)^\ell \quad (3.92)$$

$$= \frac{\tilde{\theta}_1}{z} H_\psi(z) H_\varphi(z) + \sum_{\ell=2}^{\infty} \varphi^{1-\ell} \frac{\tilde{\theta}_2^\ell \tilde{\mu}_\ell}{z^\ell} H_\psi(z)^\ell H_\varphi(z)^\ell \quad (3.93)$$

where in (3.90) we used (3.82)-(3.86) and in (3.91) we considered the following asymptotic contributions

$$E_2 = \lim_{n\to\infty} E_2^{(n)} \overset{\text{Lemma 3.2.3}}{=} \int \frac{1}{\sqrt{2\pi}} \phi(\tilde{\mu}_1 \sigma_w \sigma_x \tilde{z})^2 e^{-\frac{\tilde{z}^2}{2}} d\tilde{z} =: \tilde{\theta}_2, \quad (3.94)$$

$$E_{2k} = \lim_{n\to\infty} E_{2k}^{(n)} \overset{\text{Lemma 3.2.4}}{=} n^{1-k}\tilde{\mu}_k \left(\int \frac{\sigma_w \sigma_x}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \phi'(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1}\tilde{z}) d\tilde{z} \right)^{2k} =: n^{1-k}\tilde{\mu}_k \tilde{\theta}_2^k \tag{3.95}$$

This implies that

$$H(z) = 1 + \frac{\tilde{\theta}_1}{\varphi z} H_\psi(z) H_\varphi(z) + \sum_{\ell=2} \frac{\tilde{\theta}_2^\ell \tilde{\mu}_\ell}{\varphi^\ell z^\ell} H_\psi(z)^\ell H_\varphi(z)^\ell \tag{3.96}$$

$$= 1 + \frac{\tilde{\theta}_1 - \tilde{\mu}_1 \tilde{\theta}_2}{\varphi z} H_\psi(z) H_\varphi(z) + \sum_{\ell=1} \frac{\tilde{\mu}_\ell \tilde{\theta}_2^\ell}{\varphi^\ell z^\ell} H_\psi(z)^\ell H_\varphi(z)^\ell \tag{3.97}$$

$$= 1 + \frac{\tilde{\theta}_1 - \tilde{\mu}_1 \tilde{\theta}_2}{\varphi z} H_\psi(z) H_\varphi(z) + \frac{\varphi z}{\tilde{\theta}_2 H_\psi(z) H_\varphi(z)} \left(\sum_{\ell=0} \tilde{\mu}_\ell \left(\frac{\tilde{\theta}_2}{\varphi z} H_\psi(z) H_\varphi(z) \right)^{\ell+1} \right) - 1 \tag{3.98}$$

$$= 1 + \frac{\tilde{\theta}_1 - \tilde{\mu}_1 \tilde{\theta}_2}{\varphi z} H_\psi(z) H_\varphi(z) + \frac{\varphi z}{\tilde{\theta}_2 H_\psi(z) H_\varphi(z)} G_{\mathbf{\Sigma}} \left(\frac{\varphi z}{\tilde{\theta}_2 H_\psi(z) H_\varphi(z)} \right) - 1 \tag{3.99}$$

where we used the definition of Stieltjes Transform $G_{\mathbf{\Sigma}}(z) = \sum_{\ell=0} \frac{\tilde{\mu}_\ell}{z^{\ell+1}}$ with $\tilde{\mu}_0 = 1$. □

As a validation test, considering $\mathbf{\Sigma} = \mathbf{I}$ for which $G_{\mathbf{\Sigma}}(z) = \frac{1}{z-1}$ leads to

$$H(z) = 1 + \frac{\theta_1 - \theta_2}{\varphi z} H_\psi(z) H_\varphi(z) + \frac{\varphi z}{\theta_2 H_\psi(z) H_\varphi(z)} \frac{1}{\frac{\varphi z}{\theta_2 H_\psi(z) H_\varphi(z)} - 1} - 1 \tag{3.100}$$

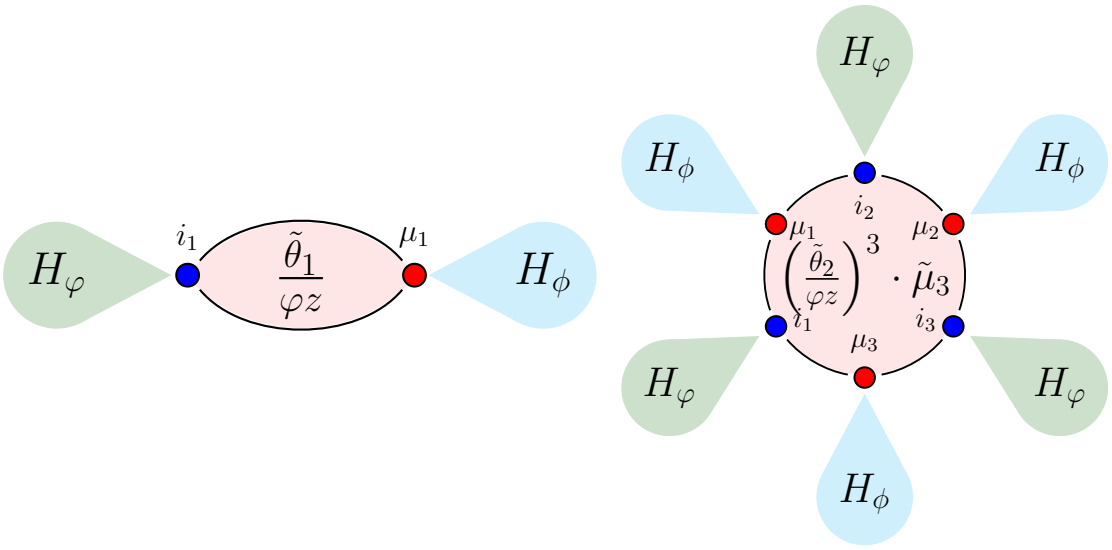
$$= 1 + \frac{\theta_1 - \theta_2}{\varphi z} H_\psi(z) H_\varphi(z) + \frac{z}{\theta_2 H_\psi(z) H_\varphi(z)} \frac{\theta_2 H_\psi(z) H_\varphi(z)}{\varphi z - \theta_2 H_\psi(z) H_\varphi(z)} - 1 \tag{3.101}$$

$$= 1 + \frac{\theta_1 - \theta_2}{\varphi z} H_\psi(z) H_\varphi(z) + \frac{\varphi z}{\varphi z - \theta_2 H_\psi(z) H_\varphi(z)} - 1 \tag{3.102}$$

$$= 1 + \frac{\theta_1 - \theta_2}{\varphi z} H_\psi(z) H_\varphi(z) + \frac{\varphi z - \varphi z + \theta_2 H_\psi(z) H_\varphi(z)}{\varphi z - \theta_2 H_\psi(z) H_\varphi(z)} \tag{3.103}$$

$$= 1 + \frac{\theta_1 - \theta_2}{\varphi z} H_\psi(z) H_\varphi(z) + \frac{\theta_2 H_\psi(z) H_\varphi(z)}{\varphi z - \theta_2 H_\psi(z) H_\varphi(z)} \tag{3.104}$$

which is the same result as in Theorem 3.2.1.



(a) $\ell = 1$ (b) $\ell = 3$

Figure 3.7: Illustration of the contributions that have to be considered for different cycles of dimension ℓ. This Figure reproduces [6, Figure 14].

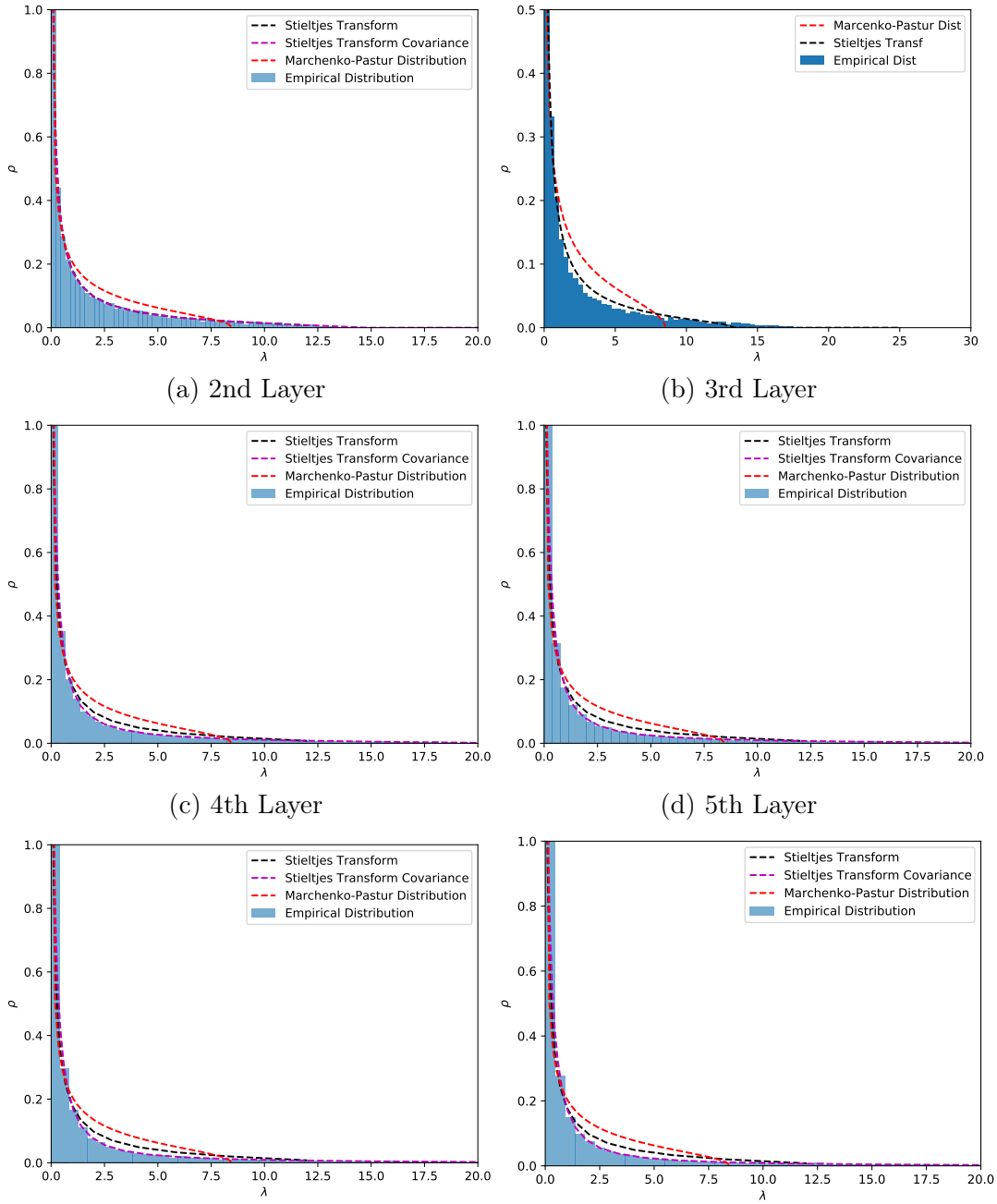
3.2.2.4 Numerical Results

In Figure 3.8 it shown how by modelling also the covariance of the input it is possible to improve the description of the eigen-distribution of the matrix $\mathbf{\Lambda}_{h^{(\ell)}}$. This is also supported by Figure 3.9 where the Wasserstein distance between the empirical distribution and the different approximation is plotted. The Stieltjes transform considering the correlation of the inputs achieves the best accuracy among the considered methods and the fact that the error does not increase with the layers, suggests that it is converging to the right solution.

3.3 Eigenspectrum of $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}^\top_{xh^{(\ell)}}\mathbf{\Sigma}_{xh^{(\ell)}}$

Given the weights $\mathbf{W}^{(\ell)}$ and $\mathbf{b}^{(\ell)}$

$$\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}^\top_{xh^{(\ell)}}\mathbf{\Sigma}_{xh^{(\ell)}} = \mathbb{E}_x[(\mathbf{W}^{(\ell)}\phi(\mathbf{h}^{(\ell)}) + \mathbf{n}^{(\ell)} + \mathbf{b}^{(\ell)})(\mathbf{W}^{(\ell)}\phi(\mathbf{h}^{(\ell)}) + \mathbf{n}^{(\ell)} + \mathbf{b}^{(\ell)})^\top] - \frac{1}{\sigma_x^2}\mathbb{E}_x[\mathbf{W}^{(\ell)}\phi(\mathbf{h}^{(\ell)})\mathbf{x}^\top]\mathbb{E}_x[\mathbf{x}\phi(\mathbf{h}^{(\ell)})^\top \mathbf{W}^{(\ell)\top}] \tag{3.105}$$



(a) 2nd Layer (b) 3rd Layer

(c) 4th Layer (d) 5th Layer

(e) 6th Layer (f) 7th Layer

Figure 3.8: Approximation of the eigenspectrum for the matrix $\mathbf{\Lambda}_{h^{(\ell)}}$ with $n = 100$ and $(\sigma_w, \sigma_b, \sigma_n) = (2, 0, 0.001)$ for six different layers. The histogram (Empirical Results) corresponds to the eigenvalues obtained on 100 different simulations and it is compared to the Marchenko-Pastur distribution with the same mean, to the distributions obtained with the Stieltjes transform of Theorem 3.2.1 (Stilejes Transform) and of Theorem 3.2.2 (Stieltjes Transform Covariance).

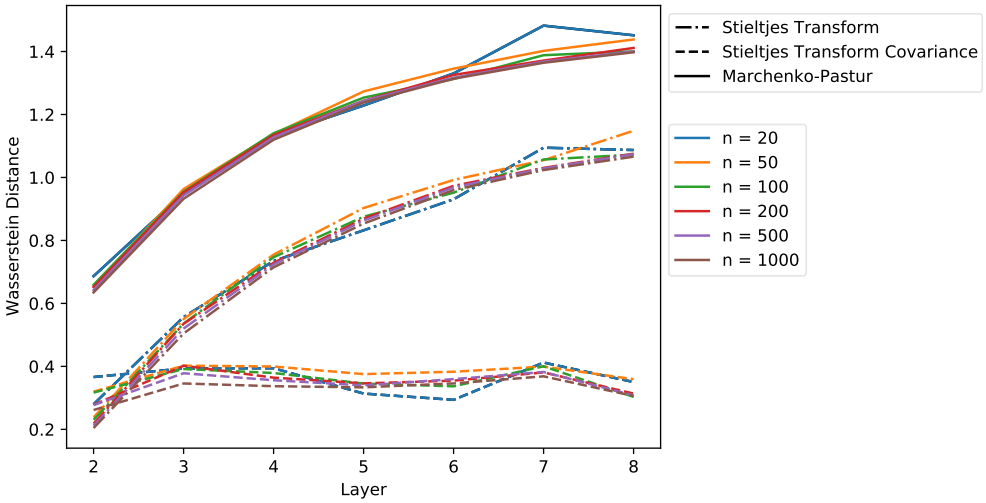


Figure 3.9: Comparison of the Wasserstein distance between the empirical distribution of $\mathbf{\Lambda}_{h^{(\ell)}}$ for different network dimensions and the distribution attained with the Stieltjes transform (dash-dotted), with the Stieltjes transform with correlated inputs (dashed) and with the Marchenko-Pastur distribution (solid). Among the approximations, the one considering the correlation of the inputs is the only to not be affected by the depth of the layer suggesting that it is able to approximate the distribution appropriately at any layer.

$$
\begin{aligned}
&= \left[\mathbf{W}^{(\ell)}, \mathbf{b}^{(\ell)} + \mathbf{n}^{(\ell)}\right] \cdot \\
&\quad \cdot \begin{bmatrix} \mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})\phi(\mathbf{h}^{(\ell)})^\top] - \frac{1}{\sigma_x^2}\mathbb{E}_x[\phi(\mathbf{h}^{(\ell)})\mathbf{x}^\top]\mathbb{E}_x[\mathbf{x}\phi(\mathbf{h}^{(\ell)})^\top] & \mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})] \\ \mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})]^\top & \mathbf{I} \end{bmatrix} \cdot \\
&\quad \cdot \left[\mathbf{W}^{(\ell)}, \mathbf{b}^{(\ell)} + \mathbf{n}^{(\ell)}\right]^\top
\end{aligned} \tag{3.106}
$$

Since only odd activation functions ϕ are considered, the vector $\mathbb{E}[\phi(\mathbf{h}^{(\ell)})]$ is going to be null and then

$$
\begin{aligned}
&\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}} \\
&\qquad = \mathbf{W}^{(\ell)} \left(\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})\phi(\mathbf{h}^{(\ell)})^\top] - \frac{1}{\sigma_x^2}\mathbb{E}_x[\phi(\mathbf{h}^{(\ell)})\mathbf{x}^\top]\mathbb{E}_x[\mathbf{x}\phi(\mathbf{h}^{(\ell)})^\top]\right) \mathbf{W}^{(\ell)^\top} \\
&\qquad\quad + (\mathbf{b}^{(\ell)} + \mathbf{n}^{(\ell)})(\mathbf{b}^{(\ell)} + \mathbf{n}^{(\ell)})^\top.
\end{aligned} \tag{3.107}
$$

The two matrices $\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})\phi(\mathbf{h}^{(\ell)})^\top]$ and $\mathbb{E}_x[\phi(\mathbf{h}^{(\ell)})\mathbf{x}^\top]\mathbb{E}_x[\mathbf{x}\phi(\mathbf{h}^{(\ell)})^\top]$ are not free relative to each other and the additivity property of the R-transform cannot

be implemented; therefore the $\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})\phi(\mathbf{h}^{(\ell)})^\top] - \mathbb{E}_x[\phi(\mathbf{h}^{(\ell)})\mathbf{x}^\top]\mathbb{E}_x[\mathbf{x}\phi(\mathbf{h}^{(\ell)})^\top]$ matrix has to be considered as one. The matrix $\mathbf{W}^{(\ell)}\mathbf{W}^{(\ell)\top}$ is free relative to $\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})\phi(\mathbf{h}^{(\ell)})^\top] - \mathbb{E}_x[\phi(\mathbf{h}^{(\ell)})\mathbf{x}^\top]\mathbb{E}_x[\mathbf{x}\phi(\mathbf{h}^{(\ell)})^\top]$, therefore, it is possible to study their Stieltjes transform thanks to the multiplicativity of the S-transform. Moreover, since the matrix $(\mathbf{b}^{(\ell)} + \mathbf{n}^{(\ell)})(\mathbf{b}^{(\ell)} + \mathbf{n}^{(\ell)})^\top$ is free relative to the matrix

$$\mathbf{W}^{(\ell)} \left(\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})\phi(\mathbf{h}^{(\ell)})^\top] - \mathbb{E}_x[\phi(\mathbf{h}^{(\ell)})\mathbf{x}^\top]\mathbb{E}_x[\mathbf{x}\phi(\mathbf{h}^{(\ell)})^\top] \right) \mathbf{W}^{(\ell)\top}$$

, then it is possible to compute the spectrum of $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$ thanks to the R-transform.

In this section we determine the Stieltjes transform of $\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})\phi(\mathbf{h}^{(\ell)})^\top] - \mathbb{E}_x[\phi(\mathbf{h}^{(\ell)})\mathbf{x}^\top]\mathbb{E}_x[\mathbf{x}\phi(\mathbf{h}^{(\ell)})^\top]$ which is central to the study of $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$. In Section 3.3.1 we propose an implicit relation related to the Stieltjes transform of the second layer. In Section 3.3.2 we expand this to deeper layers.

3.3.1 Stieltjes Transform of the Matrix $\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(1)})\phi(\mathbf{h}^{(1)})^\top] - \mathbb{E}_x[\phi(\mathbf{h}^{(1)})\mathbf{x}^\top]\mathbb{E}_x[\mathbf{x}\phi(\mathbf{h}^{(1)})^\top]$ with $\sigma_b = 0$ and $\sigma_n = 0$

To rely on the moment of methods as in Section 3.2, it is necessary to study the expectation on the input as an empirical expectation with $m \to \infty$ samples. By considering an input matrix $\mathbf{X}$, the Stieltjes transform of $\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(1)})\phi(\mathbf{h}^{(1)})^\top] - \mathbb{E}_x[\phi(\mathbf{h}^{(1)})\mathbf{x}^\top]\mathbb{E}_x[\mathbf{x}\phi(\mathbf{h}^{(1)})^\top]$ is determined by studying the eigen-distribution of

$$\mathbf{M} = \frac{1}{m}\mathbb{E}_{\mathbf{X}}\left[\phi(\mathbf{W}^{(1)}\mathbf{X})\phi(\mathbf{W}^{(1)}\mathbf{X})^\top\right] - \frac{1}{m^2\sigma_x^2}\mathbb{E}\left[\phi(\mathbf{W}^{(1)}\mathbf{X})\mathbf{X}^\top\right]\mathbb{E}\left[\phi(\mathbf{W}^{(1)}\mathbf{X})\mathbf{X}^\top\right]^\top \tag{3.108}$$

when $m \to \infty$. The definition of $\mathbf{M}$ already considers the expectation over $\mathbf{X}$ and this is necessary because of the nature of the matrix product $\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$.

Considering the admissible graphs of Definition 3.1.1, shown to be the leading order contribution also in this setting in Appendix A.2.4, we determine an implicit relation for the Stieltjes transform of $\mathbf{M}$, defined in Theorem 3.3.1. When $m/n \to \infty$ Theorem 3.3.1 holds for the matrix $\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})\phi(\mathbf{h}^{(\ell)})^\top] - \mathbb{E}_x[\phi(\mathbf{h}^{(\ell)})\mathbf{x}^\top]\mathbb{E}_x[\mathbf{x}\phi(\mathbf{h}^{(\ell)})^\top]$.

Theorem 3.3.1. *Consider the odd activation function ϕ*

$$\left| \int \frac{dz}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \phi^k(\sigma_w \sigma_x z) \right| < \infty \quad \textit{and} \quad \left| \int \frac{dz}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \phi^{(k)}(\sigma_w \sigma_x z) \right| < \infty \tag{3.109}$$

for $k > 1$ with $\phi^{(k)}$ being the k-th derivative of ϕ and the matrices $\boldsymbol{W} \in \mathbb{R}^{n_1 \times n_0}$ and $\boldsymbol{X} \in \mathbb{R}^{n_0 \times m}$ with their respective columns being sampled as follows $\boldsymbol{w}_i \sim \mathcal{N}(\boldsymbol{0}, \frac{\sigma_w^2}{n}\boldsymbol{I})$ and $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma_x^2\boldsymbol{I})$. Then, defining $\psi = \frac{n_0}{m}$ and $\varphi = \frac{n_0}{n_1}$, the matrix $\boldsymbol{M} = \frac{1}{m}\mathbb{E}_{\boldsymbol{X}}\left[\boldsymbol{Y}\boldsymbol{Y}^\top\right] - \frac{1}{\sigma_x^2 m^2}\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{Y}\boldsymbol{X}^\top\right]\mathbb{E}_{\boldsymbol{X}}\left[\boldsymbol{X}\boldsymbol{Y}^\top\right]$ with $\boldsymbol{Y} = \phi(\boldsymbol{W}\boldsymbol{X})$ has an eigen-distribution whose Stieltjes transform is asymptotically defined by

$$G_{\boldsymbol{M}}(z) \simeq \frac{1-\varphi}{z} + \frac{\varphi}{z}H(z) \tag{3.110}$$

where

$$\begin{aligned} H(z) = 1 &+ \frac{H_{\psi b}(z)H_{\varphi}(z)(\theta_1 - \sigma_x^2\sigma_w^2\theta_3^2)}{\varphi z} - \frac{H_{\psi c}(z)H_{\varphi}(z)(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)}{\varphi z} \\ &+ \frac{H_{\psi c}(z)H_{\varphi}(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)}{\varphi z - H_{\psi c}(z)H_{\varphi}(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)} \end{aligned} \tag{3.111}$$

with

$$\theta_1 = \int \frac{1}{\sqrt{2\pi}}\phi(\sigma_w\sigma_x z)^2 e^{-\frac{z^2}{2}}dz, \tag{3.112}$$

$$\theta_2 = \left(\int \frac{\sigma_w\sigma_x}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\phi'(\sigma_w\sigma_x z)dz\right)^2, \tag{3.113}$$

$$\theta_3 = \int_{z_1} \frac{\phi'(\sigma_x\sigma_w z)}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz \tag{3.114}$$

$H_{\psi\alpha}(z) = 1 + \psi\kappa_\alpha(H(z)-1)$, $H_\varphi = 1 + \varphi(H(z)-1)$, $\alpha \in \{b, c\}$, and $\kappa_b = 1 + \frac{\theta_1\sigma_x^2\sigma_w^2\theta_3^2 + \sigma_x^2\sigma_w^2\theta_3^2\theta_2 - \sigma_x^4\sigma_w^4\theta_3^4}{(\theta_1 - \sigma_x^2\sigma_w^2\theta_3^2)(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)}$, $\kappa_c = 1 + \frac{2\sigma_x^2\sigma_w^2\theta_3^2\theta_2 - \sigma_x^4\sigma_w^4\theta_3^4}{(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)^2}$.

In the following subsections, the proof of Theorem 3.3.1 is structured as for Theorem 3.2.2 by determining the moments of $\mathbf{M}$ defined according to the following expectation

$$\frac{1}{n}\mathbb{E}_{\mathbf{W}}[tr(\mathbf{M}^k)] = \frac{1}{n}\mathbb{E}_{\mathbf{W}}\left[\sum_{i_1,\ldots,i_k \in [n]} \mathbf{M}_{i_1 i_2}\mathbf{M}_{i_2 i_3}\ldots\mathbf{M}_{i_k i_1}\right] \tag{3.115}$$

$$= \frac{1}{nm^k}\mathbb{E}_{\mathbf{W}}\left[\sum_{\substack{i_1,\ldots,i_k \in [n] \\ \mu_1,\ldots,\mu_k \in [m]}} \mathbf{M}_{i_1 i_2}^{\mu_1}\mathbf{M}_{i_2 i_3}^{\mu_2}\ldots\mathbf{M}_{i_k i_1}^{\mu_k}\right] \tag{3.116}$$

where

$$\mathbf{M}_{i_1 i_2}^{\mu_1} = \mathbb{E}_{\mathbf{X}}\left[\phi(\sum_l \mathbf{W}_{i_1,l}\mathbf{X}_{l\mu_1})\phi(\sum_l \mathbf{W}_{i_2,l}\mathbf{X}_{l\mu_1})\right] - \frac{1}{m\sigma_x^2}\sum_{p,q=1}^{n,m}\mathbb{E}_{\mathbf{X}}\left[\phi(\sum_l \mathbf{W}_{i_1,l}\mathbf{X}_{l\mu_1})\mathbf{X}_{p\mu_1}\right]\mathbb{E}_{\mathbf{X}}\left[\phi(\sum_l \mathbf{W}_{i_2,l}\mathbf{X}_{lq})\mathbf{X}_{pq}\right]. \tag{3.117}$$

In Subsection 3.3.1.1 the contribution of single cycle patterns are determined and in Subsection 3.3.1.2 they are aggregated to identify the contribution to the moments of an admissible graph. This allows to define a generating function and to determine the contribution of two graphs that are connected via a vertex. These observations are then used in Subsection 3.3.1.3 to prove Theorem 3.3.1.

An interesting consequence on the Theorem 3.3.1 being applied at the first layer is reported in Corollary 3.3.1.1 whose proof is included in Appendix A.2.5.

Corollary 3.3.1.1. *For* $\boldsymbol{M} = \frac{1}{m}\mathbb{E}_{\boldsymbol{X}}\left[\boldsymbol{Y}\boldsymbol{Y}^\top\right] - \frac{1}{\sigma_x^2 m^2}\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{Y}\boldsymbol{X}^\top\right]\mathbb{E}_{\boldsymbol{X}}\left[\boldsymbol{X}\ \boldsymbol{Y}^\top\right]$ *and* $\boldsymbol{Y} = \phi(\boldsymbol{W}\boldsymbol{X})$ *as defined in Theorem 3.3.1, the eigen-distribution of the matrix* $\boldsymbol{W}\boldsymbol{M}\boldsymbol{W}^\top$ *asymptotically follows the Marchenko-Pastur distribution with mean* $\sigma_w^2(\theta_1 - \theta_2)$ *and shape* $\gamma = \frac{n_1}{n_0}$ *when* $\psi = 0$ *and* $\varphi = 1$.

3.3.1.1 $2k$-cycle Contributions

Similarly the heterogeneous input proof, the contribution of cycles of length $2k$ is first computed to then study the contribution of any admissible graph. We thus start by considering cycles of length $2k$, i.e. in Equation (3.116) $i_1 \neq i_2 \neq ... \neq i_k$ and $\mu_1 \neq \mu_2 \neq ... \neq \mu_k$.

$$\begin{aligned} E_{2k}^{(n)} &= \mathbb{E}_{\mathbf{W}}\left[\mathbf{M}_{i_1 i_2}^{\mu_1}\mathbf{M}_{i_2 i_3}^{\mu_2}...\mathbf{M}_{i_k i_1}^{\mu_k}\right] \\ &= \int_{\mathbf{W}}\left\{\left(\int \phi(\sum_l \mathbf{W}_{i_1,l}\mathbf{X}_{l\mu_1})\phi(\sum_l \mathbf{W}_{i_2,l}\mathbf{X}_{l\mu_1})\mathcal{D}\mathbf{X}\right.\right. \\ &\quad \left.\left. -\frac{1}{m\sigma_x^2}\sum_{p,q=1}^{n,m}\int \phi(\sum_l \mathbf{W}_{i_1,l}\mathbf{X}_{l\mu_1})\mathbf{X}_{p\mu_1}\mathcal{D}\mathbf{X}\int \phi(\sum_l \mathbf{W}_{i_2,l}\mathbf{X}_{lq})\mathbf{X}_{pq}\mathcal{D}\mathbf{X}\right)\right. \\ &\quad ... \end{aligned} \tag{3.118}$$

$$\left(\int \phi(\sum_l \mathbf{W}_{i_k,l}\mathbf{X}_{l,\mu_k})\phi(\sum_l \mathbf{W}_{i_1,l}\mathbf{X}_{l,\mu_k})\mathcal{D}\mathbf{X} \right.$$
$$\left. -\frac{1}{m\sigma_x^2}\sum_{p,q=1}^{n,m}\int \phi(\sum_l \mathbf{W}_{i_k,l}\mathbf{X}_{l\mu_k})\mathbf{X}_{p\mu_k}\mathcal{D}\mathbf{X}\int \phi(\sum_l \mathbf{W}_{i_1,l}\mathbf{X}_{lq})\mathbf{X}_{pq}\mathcal{D}\mathbf{X}\right)\Bigg\}\mathcal{D}\mathbf{W} \tag{3.119}$$

It is possible to reformulate the right addend thanks to the following identity

$$\int_{\mathbf{X}} \phi(\sum_{l\neq\beta}\mathbf{W}_{\alpha l}\mathbf{X}_{lp}+\mathbf{W}_{\alpha\beta}\mathbf{X}_{\beta,p})\mathbf{X}_{\beta,p}\mathcal{D}\mathbf{X} = \int_{\mathbf{X}}\sigma_x^2\phi'(\sum_l \mathbf{W}_{\alpha l}\mathbf{X}_{lp})\mathbf{W}_{\alpha\beta}\mathcal{D}\mathbf{X} \tag{3.120}$$

$$= \sigma_x^2\mathbf{W}_{\alpha\beta}\int \phi'\left(\sigma_x\sigma_w z\right)\mathcal{D}z = \sigma_x^2\mathbf{W}_{\alpha\beta}\theta_3 \tag{3.121}$$

where $\mathcal{D}z$ is the standard normal measure.

This leads to the following expression

$$E_{2k}^{(n)} = \int_{\mathbf{W}}\left\{\left(\int \phi(\sum_l \mathbf{W}_{i_1,l}\mathbf{X}_{l\mu_1})\phi(\sum_l \mathbf{W}_{i_2,l}\mathbf{X}_{l\mu_1})\mathcal{D}\mathbf{X} - \frac{\sigma_x^2}{m}\sum_{p,q=1}^{n,m}\mathbf{W}_{i_1p}\mathbf{W}_{i_2p}\theta_3^2\right)\right. \tag{3.122}$$

...

$$\left.\left(\int \phi(\sum_l \mathbf{W}_{i_k,l}\mathbf{X}_{l,\mu_k})\phi(\sum_l \mathbf{W}_{i_1,l}\mathbf{X}_{l,\mu_k})\mathcal{D}\mathbf{X} - \frac{\sigma_x^2}{m}\sum_{p,q=1}^{n,m}\mathbf{W}_{i_kp}\mathbf{W}_{i_1p}\theta_3^2\right)\right\}\mathcal{D}\mathbf{W}\mathcal{D}\mathbf{X} \tag{3.123}$$

$$= \int_{W,X}\left\{\prod_{\xi=1}^{k}\left(\phi(\sum_l \mathbf{W}_{i_\xi,l}\mathbf{X}_{l,\mu_\xi})\phi(\sum_l \mathbf{W}_{i_{\xi+1},l}\mathbf{X}_{l,\mu_\xi}) - \sigma_x^2\sum_{p=1}^{n}\mathbf{W}_{i_\xi p}\mathbf{W}_{i_{\xi+1}p}\theta_3^2\right)\right\}\mathcal{D}\mathbf{W}\mathcal{D}\mathbf{X}. \tag{3.124}$$

In the case of $k = 1$, the integral can be split in two due to its linearity, and this leads to

Lemma 3.3.2. *For* $\boldsymbol{M} = \frac{1}{m}\mathbb{E}_{\boldsymbol{X}}\left[\boldsymbol{Y}\boldsymbol{Y}^\top\right] - \frac{1}{\sigma_x^2 m^2}\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{Y}\boldsymbol{X}^\top\right]\mathbb{E}_{\boldsymbol{X}}\left[\boldsymbol{X}\ \boldsymbol{Y}^\top\right]$ *and* $\boldsymbol{Y} = \phi\left(\boldsymbol{W}\boldsymbol{X}\right)$ *as defined in Theorem 3.3.1, when* $k = 1$ *then contribution of an admissible cycle is*

$$E_2^{(n)} = \theta_1\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) - \sigma_w^2\sigma_x^2\theta_3^2 \tag{3.125}$$

with

$$\theta_1 = \int \frac{1}{\sqrt{2\pi}} \phi(\sigma_w \sigma_x z)^2 e^{-\frac{z^2}{2}} dz, \tag{3.126}$$

$$\theta_3 = \int \frac{1}{\sqrt{2\pi}} \phi' \left(\sigma_x \sigma_w z\right) e^{-\frac{z^2}{2}} dz. \tag{3.127}$$

Proof. See proof in Section A.2.1. □

However, when considering $k > 1$, we have to consider also mixed products. More complex considerations allow to determine the contribution of a general $2k$-cycle.

Lemma 3.3.3. *For* $\boldsymbol{M} = \frac{1}{m}\mathbb{E}_{\boldsymbol{X}}\left[\boldsymbol{Y}\boldsymbol{Y}^\top\right] - \frac{1}{\sigma_x^2 m^2}\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{Y}\boldsymbol{X}^\top\right]\mathbb{E}_{\boldsymbol{X}}\left[\boldsymbol{X}\ \boldsymbol{Y}^\top\right]$ *and* $\boldsymbol{Y} = \phi\left(\boldsymbol{W}\ \boldsymbol{X}\right)$ *as defined in Theorem 3.3.1, when* $k > 1$ *then contribution of an admissible cycle is*

$$E_{2k}^{(n)} = n_0^{1-k}\left(\theta_2 - \sigma_w^2 \sigma_x^2 \theta_3^2\right)^k \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \tag{3.128}$$

with

$$\theta_2 = \left(\int \frac{\sigma_w \sigma_x}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \phi'(\sigma_w \sigma_x z) z\right)^2, \tag{3.129}$$

$$\theta_3 = \int \frac{1}{\sqrt{2\pi}} \phi' \left(\sigma_x \sigma_w z\right) e^{-\frac{z^2}{2}} dz \tag{3.130}$$

Proof. See proof in Section A.2.2. □

3.3.1.2 Agglomeration of the Contributions

To compute the moments of the distribution, we follow the exact same process as in the heterogeneous case for $\boldsymbol{\Lambda}_{h^l}$ where we identify all of the possible patterns, identify their cardinality (each pattern can be achieved by selecting different vertices), and their Expected contribution thanks to the previous subsection.

To count the different patterns, after naming I_i and I_μ the identifications in Definition 3.1.1 of respectively odd and even parity, the following variable is considered similarly to [6, 72]:

Definition 3.3.1. *$\mathcal{C}(k, I_i, I_\mu, b, b^\mu, c^\mu)$ is the number of admissible pattern with $2k$ edges, I_i i-identifications, $I_\mu = b^\mu + c^\mu$ μ-identifications where c^μ are the identifications that do not define any two dimensional cycle, and with exactly b cycles of size 2, similarly to [6, 72].*

Therefore, it is possible to compute the k-th moment of the eigen-distribution

Proposition 3.3.4. *For $\boldsymbol{M} = \frac{1}{m}\mathbb{E}_{\boldsymbol{X}}\left[\boldsymbol{Y}\boldsymbol{Y}^\top\right] - \frac{1}{\sigma_x^2 m^2}\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{Y}\boldsymbol{X}^\top\right]\mathbb{E}_{\boldsymbol{X}}\left[\boldsymbol{X}\ \boldsymbol{Y}^\top\right]$ and $\boldsymbol{Y} = \phi(\boldsymbol{W}\boldsymbol{X})$ as defined in Theorem 3.3.1, the k-th moment of the eigen-distribution, for $k > 1$, is*

$$\mathfrak{m}_k^{(n)} = \mathfrak{m}_k\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \tag{3.131}$$

with

$$\begin{aligned}
\mathfrak{m}_k &= \lim_{n\to\infty} \mathfrak{m}_k^{(n)} && (3.132)\\
&= \varphi^{1-k} \sum_{I_i, I_\mu = 1}^{k} \sum_{b=0}^{I_i + I_\mu + 1} \sum_{b^\mu = 0}^{b} \frac{\mathcal{C}(k, I_i, I_\mu, b, b^\mu, c^\mu)}{2\pi\sqrt{(k-I_i)(k-I_\mu)}} (\theta_1 - \sigma_x^2\sigma_w^2\theta_3^2)^b \cdot \\
&\quad \cdot (\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)^{k-b} \kappa_b^{b^\mu} \kappa_c^{I_\mu - b^\mu} \varphi^{I_1} \psi^{I_\mu} \left(\frac{e}{k-I_i}\right)^{k-I_1} \left(\frac{e}{k-I_\mu}\right)^{k-I_\mu} && (3.133)\\
&= \varphi^{1-k} H_k && (3.134)
\end{aligned}$$

where we define $\kappa_b = 1 + \frac{\theta_1\sigma_x^2\sigma_w^2\theta_3^2 + \sigma_x^2\sigma_w^2\theta_3^2\theta_2 - \sigma_x^4\sigma_w^4\theta_3^4}{(\theta_1 - \sigma_x^2\sigma_w^2\theta_3^2)(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)}$, $\kappa_c = 1 + \frac{2\sigma_x^2\sigma_w^2\theta_3^2\theta_2 - \sigma_x^4\sigma_w^4\theta_3^4}{(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)^2}$, and based on these finally $H_k = \sum_{I_i, I_\mu = 1}^{k} \sum_{b=0}^{I_i + I_\mu + 1} \sum_{b^\mu = 0}^{b} \mathcal{C}(k, I_i, I_\mu, b, b^\mu, c^\mu)(\theta_1 - \sigma_x^2\sigma_w^2\theta_3^2)^b (\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)^{k-b} \kappa_b^{b^\mu} \kappa_c^{c^\mu} \varphi^{I_1} \phi^{I_\mu}$.

Proof. See proof in Section A.2.3. □

This allows to compute the Stieltjes transform by introducing the generating function H.

Corollary 3.3.4.1. *For $\boldsymbol{M} = \frac{1}{m}\mathbb{E}_{\boldsymbol{X}}\left[\boldsymbol{Y}\boldsymbol{Y}^\top\right] - \frac{1}{\sigma_x^2 m^2}\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{Y}\boldsymbol{X}^\top\right]\mathbb{E}_{\boldsymbol{X}}\left[\boldsymbol{X}\ \boldsymbol{Y}^\top\right]$ and $\boldsymbol{Y} = \phi(\boldsymbol{W}\boldsymbol{X})$ as defined in Theorem 3.3.1, the Stieltjes transform of the eigen-distribution of $\boldsymbol{M}$ asymptotically satisfies the following equation*

$$G_{\boldsymbol{M}}(z) \simeq \frac{1-\varphi}{z} + \frac{\varphi}{z} H(z) \tag{3.135}$$

where $H(z)$ is the generating function defined as $H(z) = 1 + \sum_{k=1}^{\infty} \frac{H_k}{z^k \varphi^k} = 1 + \frac{1}{\varphi} \sum_{k=1}^{\infty} \frac{\mathfrak{m}_k}{z^k}$.

Proof. To start with we notice that using the definition of the Stieltjes transform (3.38)

$$G_{\mathbf{M}}(z) = \sum_{k=0}^{\infty} \frac{\mathfrak{m}_k^{(n)}}{z^{k+1}} = \sum_{k=0}^{\infty} \frac{\mathfrak{m}_k}{z^{k+1}} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \tag{3.136}$$

$$= G(z) \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \tag{3.137}$$

By relying on equations (3.38) and (3.132)-(3.134)

$$G(z) = \frac{1}{z} + \sum_{k=1}^{\infty} \frac{\mathfrak{m}_k}{z^{k+1}} = \frac{1}{z} + \frac{\varphi}{z} \sum_{k=1}^{\infty} \frac{H_k}{z^k \varphi^k} \tag{3.138}$$

$$= \frac{1-\varphi}{z} + \frac{\varphi}{z} \left(1 + \sum_{k=1}^{\infty} \frac{H_k}{z^k \varphi^k}\right) = \frac{1-\varphi}{z} + \frac{\varphi}{z} H(z). \tag{3.139}$$

□

Moreover, from Proposition 3.3.4 it follows that the total contribution to the moment of a graph which is defined by the junction of two graphs, is equal to the contribution of each defining block with a penalty term that depends on the identification joining the two graphs.

Corollary 3.3.4.2. *Consider $\boldsymbol{M} = \frac{1}{m}\mathbb{E}_{\boldsymbol{X}}\left[\boldsymbol{Y}\boldsymbol{Y}^{\top}\right] - \frac{1}{\sigma_x^2 m^2}\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{Y}\boldsymbol{X}^{\top}\right]\mathbb{E}_{\boldsymbol{X}}\left[\boldsymbol{X}\ \boldsymbol{Y}^{\top}\right]$ and $\boldsymbol{Y} = \phi(\boldsymbol{W}\boldsymbol{X})$ as defined in Theorem 3.3.1, and consider a graph G_3 with $2(p+q)$ edges that is defined, via etiher an i or μ identification, by two graphs G_1 and G_2 with $2p$ and $2q$ edges. Then the asymptotic contribution, $\mathfrak{m}_{p+q}^{G_3}$, of the graph G_3 to the moment $\mathfrak{m}_{p+q}$ is equal to the product of the contributions of the block cycle graphs that define G_3 with a correction term. This is, they are joined with an i-identification*

$$\mathfrak{m}_{p+q}^{G_3} = \mathfrak{m}_p^{G_1} \frac{n_1}{n_1} \mathfrak{m}_p^{G_2} \tag{3.140}$$

and if they are joined with a μ-identification

$$\mathfrak{m}_{p+q}^{G_3} = \mathfrak{m}_p^{G_1} \frac{n_1 \kappa_\alpha}{m} \mathfrak{m}_p^{G_2} \tag{3.141}$$

where if either one of the block cycles connected to the identification are of dimension 2, then $\alpha = b$ or otherwise $\alpha = c$, and $\kappa_b = 1 + \frac{\theta_1\sigma_x^2\sigma_w^2\theta_3^2+\sigma_x^2\sigma_w^2\theta_3^2\theta_2-\sigma_x^4\sigma_w^4\theta_3^4}{(\theta_1-\sigma_x^2\sigma_w^2\theta_3^2)(\theta_2-\sigma_x^2\sigma_w^2\theta_3^2)}$, $\kappa_c = 1 + \frac{2\sigma_x^2\sigma_w^2\theta_3^2\theta_2-\sigma_x^4\sigma_w^4\theta_3^4}{(\theta_2-\sigma_x^2\sigma_w^2\theta_3^2)^2}$.

3.3.1.3 Proof of the Implicit Relation in Theorem 3.3.1

To determine the implicit relation for the Stieltjes transform exposed in Theorem 3.3.1, proven to be unique in Appendix A.2.4, a structure similar to the one in Section 3.2.2.3 is followed.

Proof. As shown in Corollary 3.3.4.2, a graph G_3, that is defined by a graph G_1 and G_2, contributes to the total moment according to the following identity.

$$\mathfrak{m}_{q+p}^{G_3} = \mathfrak{m}_q^{G_1}\frac{n_1}{n_1}\mathfrak{m}_p^{G_2} \tag{3.142}$$

when considering an i-identification or

$$\mathfrak{m}_{q+p}^{G_3} = \mathfrak{m}_q^{G_1}\frac{n_1\kappa_\alpha}{m}\mathfrak{m}_p^{G_2} \tag{3.143}$$

when considering a μ-identification where $\alpha = b$ if either of the adjacent blocks to the identifications is a 2 dimensional cycle, or $\alpha = c$ otherwise.

If a vertex v is fixed, it is possible to use the moment generating function to define the contributions of all the possible patterns that include the chosen vertex. For each of these patterns, the vertex is going to be in a cycle of dimension 2ℓ, with a contribution $\mathfrak{m}_\ell^{2\ell-cycle}$, and to each of the remaining vertices either nothing or a further graph is connected with a correction depending on the identification. By considering the limiting variables $E_{2\ell} = \lim_{n\to\infty} E_{2\ell}^{(n)}$, the total contribution, $C_{2\ell}$, to the different moments for graphs in which the vertex v is included in a 2ℓ cycle is defined by

$$C_{2\ell} = \frac{\mathfrak{m}_\ell^{2\ell-cycle}}{z^\ell}\left(1+\sum_{G_\alpha}\frac{\mathfrak{m}_{|G_\alpha|}^{G_\alpha}}{z^{|G_\alpha|}}\right)^\ell\left(1+\sum_{G_\alpha}\frac{\kappa_\alpha n_1}{m}\frac{\mathfrak{m}_{|G_\alpha|}^{G_\alpha}}{z^{|G_\alpha|}}\right)^\ell \tag{3.144}$$

$$= \frac{1}{n_1 m^\ell z^\ell}E_{2\ell}n_1^{\ell-1}m^\ell\left(1+\sum_{G_\alpha}\frac{\mathfrak{m}_{|G_\alpha|}^{G_\alpha}}{z^{|G_\alpha|}}\right)^\ell\left(1+\sum_{G_\alpha}\frac{\kappa_\alpha n_1}{m}\frac{\mathfrak{m}_{|G_\alpha|}^{G_\alpha}}{z^{|G_\alpha|}}\right)^\ell \tag{3.145}$$

$$= \frac{1}{n_1 m^\ell z^\ell} E_{2\ell} n_1^{\ell-1} m^\ell \left(1 + \frac{n_1}{n_1}\varphi(H(z)-1)\right)^\ell \left(1 + \frac{\kappa_\alpha n_1}{m}\varphi(H(z)-1)\right)^\ell \tag{3.146}$$

$$= \frac{1}{n_1 m^\ell z^\ell} E_{2\ell} n_1^{\ell-1} m^\ell \left(1 + \varphi(H(z)-1)\right)^\ell \left(1 + \kappa_\alpha \psi(H(z)-1)\right)^\ell \tag{3.147}$$

$$= \frac{1}{n_1 m^\ell z^\ell} E_{2\ell} n_1^{\ell-1} m^\ell H_\varphi^\ell H_{\psi\alpha}^\ell \tag{3.148}$$

where G_α is any possible graph, $|G_\alpha|$ is its dimension, and the following definition of the generating function was used

$$H(z) - 1 = \frac{1}{\varphi} \sum_{G_\alpha} \frac{\mathfrak{m}_{|G_\alpha|}^{G_\alpha}}{z^{|G_\alpha|}}, \tag{3.149}$$

the following variables were introduced

$$H_\varphi(z) = 1 + \varphi(H(z)-1) \quad \text{and} \quad H_{\psi\alpha}(z) = 1 + \kappa_\alpha \psi(H(z)-1), \tag{3.150}$$

and κ_α was considered to be dependent uniquely on the cycle that contains vertex v is part of. If $\ell > 1$, then $\alpha = b$ only if the block of G_α to which the cycle is attached is of dimension 2. However, the cardinality of the connection to these kinds of graphs is of a lower order than to any other kind, thus, for any cycle with $\ell > 1$ it is going to be assumed that $\alpha = c$.

Therefore, by summing the contribution $C_{2\ell}$ over all the possible fixed vertices v and over all the possible cycle dimensions ℓ, the moment generating function is retrieved.

$$\varphi(H(z)-1) = \sum_k \frac{\mathfrak{m}_k}{z^k} = \sum_{i=1}^{n_1} \sum_{\ell=1}^{\infty} C_{2\ell} \tag{3.151}$$

$$= \sum_{\ell=1}^{\infty} \frac{E_{2\ell} n_1^{\ell-1}}{z^\ell} H_{\psi\alpha}(z)^\ell H_\varphi(z)^\ell \tag{3.152}$$

$$= \frac{\theta_1 - \sigma_x^2\sigma_w^2\theta_3^2}{z} H_{\psi b}(z) H_\varphi(z) + \sum_{\ell=2}^{\infty} n_0^{1-\ell} \frac{(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)^\ell}{z^\ell} n_1^{\ell-1} H_{\psi c}(z)^\ell H_\varphi(z)^\ell \tag{3.153}$$

$$= \frac{\theta_1 - \sigma_x^2\sigma_w^2\theta_3^2}{z} H_{\psi b}(z) H_\varphi(z) + \sum_{\ell=2}^{\infty} \frac{\varphi^{1-\ell}(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)^\ell}{z^\ell} H_{\psi c}(z)^\ell H_\varphi(z)^\ell. \tag{3.154}$$

Therefore the recursive formula is

$$H(z) = 1 + \frac{H_{\psi b}(z)H_{\varphi}(z)(\theta_1 - \sigma_x^2\sigma_w^2\theta_3^2)}{\varphi z} + \sum_{q_0=2}^{\infty}\left(\frac{H_{\psi c}(z)H_{\varphi}(\theta_2 - \theta_1 - \sigma_x^2\sigma_w^2\theta_3^2)}{\varphi z}\right)^{q_0} \tag{3.155}$$

$$= 1 + \frac{H_{\psi b}(z)H_{\varphi}(z)(\theta_1 - \sigma_x^2\sigma_w^2\theta_3^2)}{\varphi z} - \frac{H_{\psi c}(z)H_{\varphi}(z)(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)}{\varphi z} + \sum_{q_0=1}^{\infty}\left(\frac{H_{\psi c}(z)H_{\varphi}(\theta_2 - \theta_3)}{\varphi z}\right)^{q_0} \tag{3.156}$$

$$= 1 + \frac{H_{\psi b}(z)H_{\varphi}(z)(\theta_1 - \sigma_x^2\sigma_w^2\theta_3^2)}{\varphi z} - \frac{H_{\psi c}(z)H_{\varphi}(z)(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)}{\varphi z} + \frac{1}{1 - \left(\frac{H_{\psi c}(z)H_{\varphi}(\theta_2 - \theta_3)}{\varphi z}\right)} - 1 \tag{3.157}$$

$$= 1 + \frac{H_{\psi b}(z)H_{\varphi}(z)(\theta_1 - \sigma_x^2\sigma_w^2\theta_3^2)}{\varphi z} - \frac{H_{\psi c}(z)H_{\varphi}(z)(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)}{\varphi z} + \frac{\varphi z}{\varphi z - H_{\psi c}(z)H_{\varphi}(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)} - 1 \tag{3.158}$$

$$= 1 + \frac{H_{\psi b}(z)H_{\varphi}(z)(\theta_1 - \sigma_x^2\sigma_w^2\theta_3^2)}{\varphi z} - \frac{H_{\psi c}(z)H_{\varphi}(z)(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)}{\varphi z} + \frac{H_{\psi c}(z)H_{\varphi}(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)}{\varphi z - H_{\psi c}(z)H_{\varphi}(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)} \tag{3.159}$$

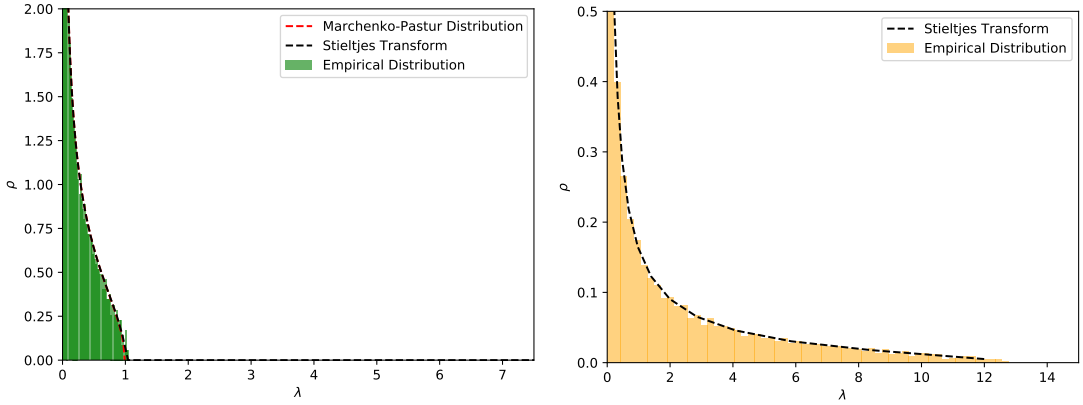
□

3.3.1.4 Numerical Experiments

In Figure 3.10 the accuracy of the predicted transform is shown for the second layer for both the matrix $\mathbf{\Lambda}_{h^{(2)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(2)}}^{\top}\mathbf{\Sigma}_{xh^{(2)}}$ and the matrix $\frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(2)}}^{\top}\mathbf{\Sigma}_{xh^{(2)}}$. It is also possible to notice how Corollary 3.3.1.1 holds in practice since the Marchenko-Pastur distribution is exactly mapped by the distribution defined in this section.

3.3.2 Stieltjes Transform of the Matrix $\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})\phi(\mathbf{h}^{(\ell)})^{\top}] - \mathbb{E}_x[\phi(\mathbf{h}^{(\ell)})\mathbf{x}^{\top}]\mathbb{E}_x[\mathbf{x}\phi(\mathbf{h}^{(\ell)})^{\top}]$ with $\sigma_b = 0$ and $\sigma_n = 0$

In this section, the matrix $\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})\phi(\mathbf{h}^{(\ell)})^{\top}] - \mathbb{E}_x[\phi(\mathbf{h}^{(\ell)})\mathbf{x}^{\top}]\mathbb{E}_x[\mathbf{x}\phi(\mathbf{h}^{(\ell)})^{\top}]$ is studied at deeper layers following a similar argument to the one for $\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(1)})\phi(\mathbf{h}^{(1)})^{\top}] -$



(a) $\mathbf{\Lambda}_{h^{(2)}} - \frac{1}{\sigma_x^{(2)}}\mathbf{\Sigma}^{\top}_{xh^{(2)}}\mathbf{\Sigma}_{xh^{(2)}}$ (b) $\frac{1}{\sigma_x^2}\mathbf{\Sigma}^{\top}_{xh^{(2)}}\mathbf{\Sigma}_{xh^{(2)}}$

Figure 3.10: Empirical distribution of the eigen-spectrum for the matrices $\mathbf{\Lambda}_{h^{(2)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}^{\top}_{xh^{(2)}}\mathbf{\Sigma}_{xh^{(2)}}$ and $\frac{1}{\sigma_x^2}\mathbf{\Sigma}^{\top}_{xh^{(2)}}\mathbf{\Sigma}_{xh^{(2)}}$ with $n = 100$ and $(\sigma_w, \sigma_b, \sigma_n) = (2, 0, 0.001)$.

$\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(1)})\mathbf{x}^{\top}]\mathbb{E}_{\mathbf{x}}[\mathbf{x}\phi(\mathbf{h}^{(1)})^{\top}]$. Therefore we study the following matrix

$$\mathbf{M} = \frac{1}{m}\mathbb{E}_{\mathbf{X}}\left[\phi(\mathbf{W}^{(\ell)}\mathbf{Y}^{(\ell)})\phi(\mathbf{W}^{(\ell)}\mathbf{Y}^{(\ell)})^{\top}\right] - \frac{1}{m^2\sigma_x^2}\mathbb{E}_{\mathbf{x}}\left[\phi(\mathbf{W}^{(\ell)}\mathbf{Y}^{(\ell)})\mathbf{X}^{\top}\right]\mathbb{E}_{\mathbf{x}}\left[\phi(\mathbf{W}^{(\ell)}\mathbf{Y}^{(\ell)})\mathbf{X}^{\top}\right]. \tag{3.160}$$

The consequent Stieltjes transform introduced in Theorem 3.3.5 however has to rely on the assumption that all the elements in the hidden layer are independent since it is not amenable keeping track of the covariance through the layers, as shown in Lemma 3.3.7.

Theorem 3.3.5. *Consider the odd activation function* ϕ

$$\left|\int \frac{dz}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\phi^k(\sigma_w\sigma_x z)\right| < \infty \quad \textit{and} \quad \left|\int \frac{dz}{\sqrt{2\pi}} e^{-\frac{z^2}{2}}\phi^{(k)}(\sigma_w\sigma_x z)\right| < \infty \tag{3.161}$$

for $k > 1$ *with* $\phi^{(k)}$ *being the* k*-th derivative of* ϕ*, for each layer the matrices* $\boldsymbol{W}^{(\ell)} \in \mathbb{R}^{n_\ell \times n_{\ell-1}}$*, an input matrix* $\boldsymbol{X} \in \mathbb{R}^{n_0 \times m}$ *with their respective columns being sampled as follows* $\boldsymbol{w}_i^{(\ell)} \sim \mathcal{N}(\boldsymbol{0}, \frac{\sigma_w^2}{n_{\ell-1}}\boldsymbol{I})$ *and* $\boldsymbol{x}_i \sim \mathcal{N}(\boldsymbol{0}, \sigma_x^2\boldsymbol{I})$*. Then define* $\boldsymbol{H}_{i,p}^{(\ell)} = \sum_{k_\ell=1}^{n} \boldsymbol{W}_{i,k_\ell}^{(\ell)}\boldsymbol{Y}_{k_\ell,p}^{(\ell)}$ *with* $\boldsymbol{Y}_{i,p}^{(\ell)} = \phi(\boldsymbol{H}_{i,p}^{(\ell-1)})$ *and with* $\boldsymbol{Y}^{(1)} = \boldsymbol{X}$*,* $\psi = \frac{n_0}{m}$*, and* $\varphi = \frac{n_0}{n_1}$*. Under the assumption that each column in* $\boldsymbol{Y}^{(\ell)}$ *is distributed according to* $\boldsymbol{Y}_{:p}^{(\ell)} \sim \mathcal{N}(\boldsymbol{0}, q^{(\ell)}\boldsymbol{I})$*, the matrix*

$$\boldsymbol{M} = \frac{1}{m}\mathbb{E}_{\boldsymbol{X}}\left[\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})^{\top}\right]$$

$$- \frac{1}{m^2\sigma_x^2}\mathbb{E}_{\boldsymbol{x}}\left[\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})\boldsymbol{X}^\top\right]\mathbb{E}_{\boldsymbol{x}}\left[\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})\boldsymbol{X}^\top\right] \tag{3.162}$$

has an eigen-distribution whose Stieltjes transform is asymptotically defined by

$$G_{\boldsymbol{M}}(z) \simeq \frac{1-\varphi}{z} + \frac{\varphi}{z}H(z) \tag{3.163}$$

where

$$\begin{aligned} H(z) =& 1 + \frac{H_{\psi b}(z)H_{\varphi}(z)\left(\theta_1^{(\ell)} - \sigma_x^2\sigma_w^{2\ell}\left(\prod_{l=1}^{\ell}\theta_3^{(l)}\right)^2\right)}{\varphi z} \\ &- \frac{H_{\psi c}(z)H_{\varphi}(z)\left(\theta_2^{(\ell)} - \sigma_x^2\sigma_w^{2\ell}\left(\prod_{l=1}^{\ell}\theta_3^{(l)}\right)^2\right)}{\varphi z} \\ &+ \frac{H_{\psi c}(z)H_{\varphi}\left(\theta_2^{(\ell)} - \sigma_x^2\sigma_w^{2\ell}\left(\prod_{l=1}^{\ell}\theta_3^{(l)}\right)^2\right)}{\varphi z - H_{\psi c}(z)H_{\varphi}\left(\theta_2^{(\ell)} - \sigma_x^2\sigma_w^{2\ell}\left(\prod_{l=1}^{\ell}\theta_3^{(l)}\right)^2\right)} \end{aligned} \tag{3.164}$$

with

$$\theta_1^{(\ell)} = \int \frac{1}{\sqrt{2\pi}}\phi(\sigma_w\sqrt{q^{(\ell)}}z)^2 e^{-\frac{z^2}{2}}dz, \tag{3.165}$$

$$\theta_2^{(\ell)} = \left(\int \frac{\sqrt{q^{(\ell)}}}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\phi'(\sqrt{q^{(\ell)}}z)dz\right)^2, \tag{3.166}$$

$$\theta_3^{(\ell)} = \int \frac{\phi'\left(\sqrt{q^{(\ell)}}z\right)}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}dz \tag{3.167}$$

$H_{\psi\alpha}(z) = 1 + \kappa_\alpha\psi(H(z)-1)$, $H_\varphi = 1 + \varphi(H(z)-1)$, $\alpha \in \{b, c\}$, *and* $\kappa_b = 1 + \frac{\theta_1\sigma_x^2\sigma_w^2\theta_3^2 + \sigma_x^2\sigma_w^2\theta_3^2\theta_2 - \sigma_x^4\sigma_w^4\theta_3^4}{(\theta_1 - \sigma_x^2\sigma_w^2\theta_3^2)(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)}$, $\kappa_c = 1 + \frac{2\sigma_x^2\sigma_w^2\theta_3^2\theta_2 - \sigma_x^4\sigma_w^4\theta_3^4}{(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)^2}$.

The proof of Theorem 3.3.5 relies on the proof of Theorem 3.3.1 to implement the moments method. This subsection will focus on implementing the method of moments by considering the following expectation

$$\frac{1}{n}\mathbb{E}_{\mathbf{W}}[tr(\mathbf{M}^k)] = \frac{1}{n}\mathbb{E}_{\mathbf{W}}\left[\sum_{i_1,\ldots,i_k\in[n]}\mathbf{M}_{i_1i_2}\mathbf{M}_{i_2i_3}\ldots\mathbf{M}_{i_ki_1}\right] \tag{3.168}$$

$$= \frac{1}{nm^k}\mathbb{E}_{\mathbf{W}}\left[\sum_{\substack{i_1,\dots,i_k\in[n]\\ \mu_1,\dots,\mu_k\in[m]}} \mathbf{M}^{\mu_1}_{i_1i_2}\mathbf{M}^{\mu_2}_{i_2i_3}...\mathbf{M}^{\mu_k}_{i_ki_1}\right] \tag{3.169}$$

where

$$\begin{aligned}\mathbf{M}^{\mu_1}_{i_1i_2} =&\mathbb{E}_{\mathbf{X}}\left[\phi(\sum_l \mathbf{W}^{(\ell)}_{i_1,l}\mathbf{Y}^{(\ell)}_{l\mu_1})\phi(\sum_l \mathbf{W}^{(\ell)}_{i_2,l}\mathbf{Y}^{(\ell)}_{l\mu_1})\right]\\ &-\frac{1}{m\sigma_x^2}\sum_{p,q=1}^{n,m}\mathbb{E}_{\mathbf{X}}\left[\phi(\sum_l \mathbf{W}^{(\ell)}_{i_1,l}\mathbf{Y}^{(\ell)}_{l\mu_1})\mathbf{X}_{p\mu_1}\right]\mathbb{E}_{\mathbf{X}}\left[\phi(\sum_l \mathbf{W}^{(\ell)}_{i_2,l}\mathbf{Y}^{(\ell)}_{lq})\mathbf{X}_{pq}\right].\end{aligned} \tag{3.170}$$

In Section 3.3.2.1 the contribution of simple cycles is computed and it is shown how it is necessary to include the assumption of the elements in the hidden layer being independent, i.e. $\mathbf{Y}^{(\ell)}_{:p} \sim \mathcal{N}(\mathbf{0}, q^{(\ell)}\mathbf{I})$, to retrieve a tractable analytical formulation. Once this assumption is considered, then in Section 3.3.2.2 two updated Lemmas from Section 3.3.1.2 are presented and their expression allows to state that the proof of Theorem 3.3.5 is the same as for Theorem 3.3.1. The results in Section 3.3.2.2 are not proven explicitly since their proofs trivially consist in renaming some variable in the proofs for Theorem 3.3.1 .

3.3.2.1 $2k$-cycle Contributions

The contribution of cycles of length $2k$ is first computed to then study the contribution of any admissible graph. We then consider cycles of length $2k$, i.e. in Equation (3.169) $i_1 \neq i_2 \neq ... \neq i_k$ and $\mu_1 \neq \mu_2 \neq ... \neq \mu_k$

$$\begin{aligned}E^{(\ell,n)}_{2k} &= \mathbb{E}_{\mathbf{W}}\left[\mathbf{M}^{\mu_1}_{i_1i_2}\mathbf{M}^{\mu_2}_{i_2i_3}...\mathbf{M}^{\mu_k}_{i_ki_1}\right] \\ &= \int_{\mathbf{W}}\left\{\left(\int \phi(\sum_l \mathbf{W}^{(\ell)}_{i_1,l}\mathbf{Y}^{(\ell)}_{l\mu_1})\phi(\sum_l \mathbf{W}^{(\ell)}_{i_2,l}\mathbf{Y}^{(\ell)}_{l\mu_1})\mathcal{D}\mathbf{X}\right.\right.\\ &\quad\left.\left.-\frac{1}{m\sigma_x^2}\sum_{p,q=1}^{n,m}\int \phi(\sum_l \mathbf{W}^{(\ell)}_{i_1,l}\mathbf{Y}^{(\ell)}_{l\mu_1})\mathbf{X}_{p\mu_1}\mathcal{D}\mathbf{X}\int \phi(\sum_l \mathbf{W}^{(\ell)}_{i_2,l}\mathbf{Y}^{(\ell)}_{lq})\mathbf{X}_{pq}\mathcal{D}\mathbf{X}\right)\right.\\ &\quad ...\\ &\quad\left(\int \phi(\sum_l \mathbf{W}^{(\ell)}_{i_k,l}\mathbf{Y}^{(\ell)}_{l\mu_k})\phi(\sum_l \mathbf{W}^{(\ell)}_{i_1,l}\mathbf{Y}^{(\ell)}_{l\mu_k})\mathcal{D}\mathbf{X}\right.\end{aligned} \tag{3.171}$$

$$-\frac{1}{m\sigma_x^2}\sum_{p,q=1}^{n,m}\int\phi(\sum_l \mathbf{W}_{i_k,l}^{(\ell)}\mathbf{Y}_{l\mu_k}^{(\ell)})\mathbf{X}_{p\mu_k}\mathcal{D}\mathbf{X}\int\phi(\sum_l \mathbf{W}_{i_1,l}^{(\ell)}\mathbf{Y}_{lq}^{(\ell)})\mathbf{X}_{pq}\mathcal{D}\mathbf{X}\Bigg)\Bigg\}\mathcal{D}\mathbf{W} \tag{3.172}$$

This can be rewritten thanks to the following identity.

$$\int_{\mathbf{X}}\phi(\mathbf{H}_{\alpha p}^{(\ell)})\mathbf{X}_{\beta p}\mathcal{D}\mathbf{X}_{\beta p} = \int_{\mathbf{X}}\phi(\mathbf{H}_{\alpha p}^{(\ell)})\mathbf{X}_{\beta p}e^{-\frac{\mathbf{x}_{\beta p}^2}{2\sigma_x^2}}d\mathbf{X} \tag{3.173}$$

$$= -\sigma_x^2\left[\phi(\mathbf{H}_{\alpha p}^{(\ell)})\mathbf{X}_{\beta p}e^{-\frac{\mathbf{x}_{\beta p}^2}{2\sigma_x^2}}\right]_{-\infty}^{+\infty} + \sigma_x^2\int_{\mathbf{X}}\frac{\partial\left(\phi(\mathbf{H}_{\alpha p}^{(\ell)})\right)}{\partial\mathbf{X}_{\beta p}}e^{-\frac{\mathbf{x}_{\beta p}^2}{2\sigma_x^2}}d\mathbf{X}_{\beta p} \tag{3.174}$$

$$= \sigma_x^2\int_{\mathbf{X}}\phi'(\mathbf{H}_{\alpha p}^{(\ell)})\sum_{k_\ell}\mathbf{W}_{\alpha k_\ell}^{(\ell)}\frac{\partial\left(\phi(\mathbf{H}_{k_\ell p}^{(\ell-1)})\right)}{\partial\mathbf{X}_{\beta p}}\mathcal{D}\mathbf{X} \tag{3.175}$$

$$= \sigma_x^2\int_{\mathcal{X}}\phi'(\mathbf{H}_{k_\ell p}^{(\ell)})\sum_{k_\ell}\mathbf{W}_{\alpha k_\ell}^{(\ell)}\left(\prod_{j=1}^{\ell-2}\left(\phi'(\mathbf{H}_{k_{\ell-j+1}p}^{(\ell-j)})\sum_{k_{\ell-j}}\mathbf{W}_{k_{\ell-j+1}k_{\ell-j}}^{(\ell-j)}\right)\right)\frac{\partial\left(\phi(\mathbf{H}_{\alpha p}^{(1)})\right)}{\partial\mathbf{X}_{\beta p}}\mathcal{D}\mathbf{X} \tag{3.176}$$

$$= \sigma_x^2\int_{\mathbf{X}}\phi'(\mathbf{H}_{k_\ell p}^{(\ell)})\sum_{k_\ell}\mathbf{W}_{\alpha k_\ell}^{(\ell)}\left(\prod_{j=1}^{\ell-2}\left(\phi'(\mathbf{H}_{k_{\ell-j+1}p}^{(\ell-j)})\sum_{k_{\ell-j}}\mathbf{W}_{k_{\ell-j+1}k_{\ell-j}}^{(\ell-j)}\right)\right)\phi'(\mathbf{H}_{k_1 p}^{(1)})\mathbf{W}_{k_2\beta}^{(1)}\mathcal{D}\mathbf{X} \tag{3.177}$$

$$= \sigma_x^2\sum_{k_\ell}\mathbf{W}_{\alpha k_\ell}^{(\ell)}\left(\prod_{j=1}^{\ell-2}\left(\sum_{k_{\ell-j}}\mathbf{W}_{k_{\ell-j+1}k_{\ell-j}}^{(\ell-j)}\right)\right)\mathbf{W}_{k_2\beta}^{(1)}\int_{\mathbf{X}}\left(\prod_{j=1}^{\ell}\phi'(\mathbf{H}_{k_{\ell-j+1}p}^{(\ell-j)})\right)\mathcal{D}\mathbf{X} \tag{3.178}$$

$$= \sigma_x^2\mathfrak{W}_{\alpha\beta}^{(\ell)}\int_{\mathbf{X}}\left(\prod_{j=1}^{\ell}\phi'(\mathbf{H}_{k_{\ell-j+1}p}^{(\ell-j)})\right)\mathcal{D}\mathbf{X} \tag{3.179}$$

where the new variable $\mathfrak{W}_{\alpha\beta}^{(\ell)} = \sum_{k_\ell}\mathbf{W}_{\alpha k_\ell}^{(\ell)}\left(\prod_{j=1}^{\ell-2}\left(\sum_{k_{\ell-j}}\mathbf{W}_{k_{\ell-j+1}k_{\ell-j}}^{(\ell-j)}\right)\right)\mathbf{W}_{k_2\beta}^{(1)} = \sum_{k_\ell}\mathbf{W}_{\alpha k_\ell}^{(\ell)}\mathfrak{W}_{k_\ell\beta}^{(\ell-1)}$ is introduced and the symmetry of the problem for which $\mathbf{H}_{\alpha\beta}^{(\ell+1)} \sim \mathbf{H}_{\gamma\delta}^{(\ell)}$ for any α, β, γ, and δ, is used to isolate the terms in $\int_{\mathbf{X}}\left(\prod_{j=1}^{\ell}\phi'(\mathbf{H}_{k_{\ell-j+1}p}^{(\ell-j)})\right)\mathcal{D}\mathbf{X}$.

Therefore, by defining

$$\theta_3^{(\ell)} = \int_{\mathbf{X}}\phi'(\mathbf{H}_{\alpha\beta}^{(\ell)})\mathcal{D}\mathbf{X} = \int\phi'(\sqrt{q^{(\ell)}}z\mathcal{D}z)$$

we can consider

$$\mathbf{M}_{i_1 i_2}^{\mu_1} = \int \phi(\sum_l \mathbf{W}_{i_1,l}^{(\ell)} \mathbf{Y}_{l\mu_1}^{(\ell)}) \phi(\sum_l \mathbf{W}_{i_2,l}^{(\ell)} \mathbf{Y}_{l\mu_1}^{(\ell)}) \mathcal{D}\mathbf{X} - \frac{1}{m\sigma_x^2} \sum_{p,q=1}^{n,m} \int \phi(\sum_l \mathbf{W}_{i_1,l}^{(\ell)} \mathbf{Y}_{l\mu_1}^{(\ell)}) \mathbf{X}_{p\mu_1} \mathcal{D}\mathbf{X} \int \phi(\sum_l \mathbf{W}_{i_2,l}^{(\ell)} \mathbf{Y}_{lq}^{(\ell)}) \mathbf{X}_{pq} \mathcal{D}\mathbf{X} \tag{3.180}$$

$$= \int \phi(\sum_l \mathbf{W}_{i_1,l}^{(\ell)} \mathbf{Y}_{l\mu_1}^{(\ell)}) \phi(\sum_l \mathbf{W}_{i_2,l}^{(\ell)} \mathbf{Y}_{l\mu_1}^{(\ell)}) \mathcal{D}\mathbf{X} - \frac{\sigma_x^2}{m} \sum_{p,q=1}^{n,m} \mathfrak{W}_{i_1 p}^{(\ell)} \int_{\mathbf{X}} \left(\prod_{j=1}^{\ell} \phi'(\mathbf{H}_{k_\ell \mu_1}^{(\ell-1)}) \right) \mathcal{D}\mathbf{X} \mathfrak{W}_{i_2 p}^{(\ell)} \int_{\mathbf{X}} \left(\prod_{j=1}^{\ell} \phi'(\mathbf{H}_{k_\ell q}^{(\ell-1)}) \right) \mathcal{D}\mathbf{X} \tag{3.181}$$

$$= \int \phi(\sum_l \mathbf{W}_{i_1,l}^{(\ell)} \mathbf{Y}_{l\mu_1}^{(\ell)}) \phi(\sum_l \mathbf{W}_{i_2,l}^{(\ell)} \mathbf{Y}_{l\mu_1}^{(\ell)}) \mathcal{D}\mathbf{X} - \sigma_x^2 \sum_p^n \mathfrak{W}_{i_1 p}^{(\ell)} \mathfrak{W}_{i_2 p}^{(\ell)} \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right)^2 \tag{3.182}$$

and we therefore aim at computing

$$E_{2k}^{(\ell,n)} = \int_{\mathbf{WX}} \left\{ \prod_{\xi=1}^{k} \left(\phi(\sum_l \mathbf{W}_{i_\xi,l}^{(\ell)} \mathbf{Y}_{l\mu_\xi}^{(\ell)}) \phi(\sum_l \mathbf{W}_{i_{\xi+1},l}^{(\ell)} \mathbf{Y}_{l\mu_\xi}^{(\ell)}) - \sigma_x^2 \sum_p^n \mathfrak{W}_{i_\xi p}^{(\ell)} \mathfrak{W}_{i_{\xi+1} p}^{(\ell)} \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right)^2 \right) \right\} \mathcal{D}\mathbf{X} \mathcal{D}\mathbf{W} \tag{3.183}$$

$$= \sum_{n_w=0}^{k} \binom{k}{n_w} \sum_{\substack{\mathbf{n}_\phi \ s.t. \\ \|\mathbf{n}_\phi\|_1 = n - n_w}} \sum_{p=1}^{n} E_w^{(k,n_w,\mathbf{n}_\phi,p)} \tag{3.184}$$

where n_w is the number of terms of the type $\mathfrak{W}_{i_\xi p}^{(\ell)} \mathfrak{W}_{i_{\xi+1} p}^{(\ell)} \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right)^2$ in one addend of the expansion of the product in $E_{2k}^{(\ell,n)}$, and $\mathbf{n}_\phi$ corresponds to a vector whose entries define how many terms of the type $\phi(\sum_l \mathbf{W}_{i_\xi,l}^{(\ell)} \mathbf{Y}_{l\mu_\xi}^{(\ell)}) \phi(\sum_l \mathbf{W}_{i_{\xi+1},l}^{(\ell)} \mathbf{Y}_{l\mu_\xi}^{(\ell)})$ are sequentially included between two terms of the type $\mathfrak{W}_{i_\xi p}^{(\ell)} \mathfrak{W}_{i_{\xi+1} p}^{(\ell)} \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right)^2$.

To achieve accurate modelling of the contribution of $E_{2k}^{(\ell,n)}$ for deeper layers, it is necessary to to keep into consideration the covariance matrix of each hidden layer $\mathbf{Y}^{(\ell)}$. Therefore, at each layer the matrix $\tilde{\mathbf{W}}^{(\ell)} \mathbf{\Sigma}^{(\ell)^{1/2}} \tilde{\mathbf{Y}}$ is going to be considered

instead of $\mathbf{W}^{(\ell)}\mathbf{Y}^{(\ell)}$. For $k = 1$, the contribution of a two dimensional cycle is determined as follows

Lemma 3.3.6. *For the matrix*

$$\boldsymbol{M} = \frac{1}{m}\mathbb{E}_{\boldsymbol{X}}\left[\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})^{\top}\right] - \frac{1}{m^2\sigma_x^2}\mathbb{E}_{\boldsymbol{x}}\left[\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})\boldsymbol{X}^{\top}\right]\mathbb{E}_{\boldsymbol{x}}\left[\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})\boldsymbol{X}^{\top}\right] \tag{3.185}$$

and $\boldsymbol{Y}^{(\ell)} = \phi\left(\boldsymbol{H}^{(\ell-1)}\right)$ *as defined in Theorem 3.3.5 without the hypothesis of independence for the elements of* $\boldsymbol{Y}^{(\ell)}$*, i.e.* $\boldsymbol{Y}^{(\ell)}_{:p} \sim \mathcal{N}(\boldsymbol{0}, q^{(\ell)}\boldsymbol{\Sigma}^{(\ell)})$*, when* $k = 1$ *then the contribution of an admissible cycle is*

$$E_2^{(\ell,n)} = \int \tilde{\theta}_1^{(\ell,n)}\mathcal{D}\boldsymbol{\Sigma}^{(\ell)}\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) - \sigma_x^2\sigma_w^{2\ell}\prod_{l=1}^{\ell}(\theta_3^{(l)})^2 \tag{3.186}$$

where

$$\tilde{\theta}_1^{(\ell,n)} = \int \frac{1}{\sqrt{2\pi}}\phi(\sqrt{q^{(\ell)}}\sqrt{\frac{tr(\boldsymbol{\Sigma}^{(\ell)})}{n}}z)^2 e^{-\frac{z^2}{2}}dz \tag{3.187}$$

$$\theta_3^{(\ell)} = \int \phi'(\sqrt{q^{(\ell)}}z)e^{-\frac{z^2}{2}}dz. \tag{3.188}$$

Proof. See proof in Section A.3.1. □

For $k > 1$, this modelling leads to the following result for an addend ω in the expansion of (3.184) with $\mathbf{n}_\phi$ and n_w

Lemma 3.3.7. *For the matrix*

$$\boldsymbol{M} = \frac{1}{m}\mathbb{E}_{\boldsymbol{X}}\left[\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})^{\top}\right] - \frac{1}{m^2\sigma_x^2}\mathbb{E}_{\boldsymbol{x}}\left[\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})\boldsymbol{X}^{\top}\right]\mathbb{E}_{\boldsymbol{x}}\left[\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})\boldsymbol{X}^{\top}\right] \tag{3.189}$$

and $\boldsymbol{Y}^{(\ell)} = \phi\left(\boldsymbol{H}^{(\ell-1)}\right)$ *as defined in Theorem 3.3.5 without the hypothesis of independence for the elements of* $\boldsymbol{Y}^{(\ell)}$*, i.e.* $\boldsymbol{Y}^{(\ell)}_{:p} \sim \mathcal{N}(\boldsymbol{0}, q^{(\ell)}\boldsymbol{\Sigma}^{(\ell)})$*, when* $k > 1$ *then*

the contribution of a single term in the expanded sum of (3.184) *is*

$$E_{\omega}^{(n,k,n_w,\boldsymbol{n}_\phi,p)} = \int \left(-\frac{\sigma_w^2\sigma_x^2\left(\prod_{l=1}^{\ell}\theta_3^{(l)}\right)^2}{n} \right)^{n_w} \left(\frac{\tilde{\theta}_2^{(\ell,n)}}{n} \right)^{k-n_w} \left(\prod_{\xi=1} \mu_{\boldsymbol{n}_\phi^{(\xi)}}^{(n)} \right) \tag{3.190}$$

where $\mu_k^{(n)} = \frac{tr(\boldsymbol{\Sigma}^{(\ell)^k})}{n}$ *and*

$$\tilde{\theta}_2^{(\ell,n)} = \left(\int \frac{\sqrt{q^{(\ell)}}}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \phi'(\sqrt{q^{(\ell)}}\sqrt{\frac{tr(\boldsymbol{\Sigma}^{(\ell)})}{n}}\tilde{z}) d\tilde{z} \right)^2, \tag{3.191}$$

$$\theta_3^{(\ell)} = \int \phi'(\sqrt{q^{(\ell)}}z) e^{-\frac{z^2}{2}} dz. \tag{3.192}$$

Proof. See proof in Section A.3.2. □

The dependence of $E_{\omega}^{(n,k,n_w,n_\phi,p)}$ on the moments of $\boldsymbol{\Sigma}^{(\ell)}$ is such that it is not possible to aggregate explicitly all of the contributions in (3.184). For this reason, it is necessary to introduce a new assumption that $\boldsymbol{\Sigma} = \mathbf{I}$. With this new assumption, it is then possible to compute explicitly the contribution of cycles with $k > 1$.

Lemma 3.3.8. *For the matrix*

$$\boldsymbol{M} = \frac{1}{m}\mathbb{E}_{\boldsymbol{X}}\left[\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})^\top\right] - \frac{1}{m^2\sigma_x^2}\mathbb{E}_{\boldsymbol{x}}\left[\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})\boldsymbol{X}^\top\right]\mathbb{E}_{\boldsymbol{x}}\left[\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})\boldsymbol{X}^\top\right] \tag{3.193}$$

and $\boldsymbol{Y}^{(\ell)} = \phi\left(\boldsymbol{H}^{(\ell-1)}\right)$ *as defined in Theorem 3.3.5, i.e.* $\boldsymbol{Y}_{:p}^{(\ell)} \sim \mathcal{N}(\boldsymbol{0}, q^{(\ell)}\boldsymbol{I})$*, when* $k > 1$ *then contribution of an admissible cycle is*

$$E_{2k} = n_0^{1-k}\left(\theta_2^{(\ell)} - \sigma_w^{2\ell}\sigma_x^2\left(\prod_{l=1}^{\ell}\theta_3^{(\ell)}\right)^2\right)^k\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \tag{3.194}$$

where

$$\theta_2^{(\ell)} = \left(\int \frac{\sqrt{q^{(\ell)}}}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \phi'(\sqrt{q^{(\ell)}}\tilde{z}) d\tilde{z} \right)^2, \tag{3.195}$$

$$\theta_3^{(\ell)} = \int \phi'(\sqrt{q^{(\ell)}}z) e^{-\frac{z^2}{2}} dz. \tag{3.196}$$

Proof. See proof in Section A.3.3. □

When considering $\mathbf{\Sigma}^{(\ell)} = \mathbf{I}$ then Lemma 3.3.6 is updated as follows

Lemma 3.3.9. *For the matrix*

$$\boldsymbol{M} = \frac{1}{m}\mathbb{E}_{\boldsymbol{X}}\left[\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})^\top\right] - \frac{1}{m^2\sigma_x^2}\mathbb{E}_{\boldsymbol{x}}\left[\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})\boldsymbol{X}^\top\right]\mathbb{E}_{\boldsymbol{x}}\left[\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})\boldsymbol{X}^\top\right] \tag{3.197}$$

and $\boldsymbol{Y}^{(\ell)} = \phi\left(\boldsymbol{H}^{(\ell-1)}\right)$ *as defined in Theorem 3.3.5 without the hypothesis of independence for the elements of* $\boldsymbol{Y}^{(\ell)}$*, i.e.* $\boldsymbol{Y}^{(\ell)}_{:p} \sim \mathcal{N}(\boldsymbol{0}, q^{(\ell)}\boldsymbol{I})$*, when* $k = 1$ *then the contribution of an admissible cycle is*

$$E_2^{(\ell,n)} = \theta_1^{(\ell)}\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) - \sigma_x^2\sigma_w^{2\ell}\prod_{l=1}^{\ell}(\theta_3^{(l)})^2 \tag{3.198}$$

where

$$\theta_1^{(\ell)} = \int \frac{1}{\sqrt{2\pi}}\phi(\sqrt{q^{(\ell)}}z)^2 e^{-\frac{z^2}{2}}dz \tag{3.199}$$

$$\theta_3^{(\ell)} = \int \phi'(\sqrt{q^{(\ell)}}z)e^{-\frac{z^2}{2}}dz. \tag{3.200}$$

3.3.2.2 Aggregation of the Contributions and Implicit Relation in Theorem 3.3.5

Once the hypothesis of independence of the elements in the hidden layers is considered, the propositions for the aggregation of the contributions are achieved in the same exact way as for the first layer of $\mathbf{\Lambda}_{h^{(1)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}^\top_{xh^{(1)}}\mathbf{\Sigma}_{xh^{(1)}}$, leading to the following two Lemmas.

Lemma 3.3.10. *For the matrix*

$$\boldsymbol{M} = \frac{1}{m}\mathbb{E}_{\boldsymbol{X}}\left[\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})^\top\right] - \frac{1}{m^2\sigma_x^2}\mathbb{E}_{\boldsymbol{x}}\left[\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})\boldsymbol{X}^\top\right]\mathbb{E}_{\boldsymbol{x}}\left[\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})\boldsymbol{X}^\top\right] \tag{3.201}$$

and $\boldsymbol{Y}^{(\ell)} = \phi\left(\boldsymbol{H}^{(\ell-1)}\right)$ *as defined in Theorem 3.3.5, the Stieltjes transform of the eigen-distribution of* $\boldsymbol{M}$ *asymptotically satisfies the following equation*

$$G_{\boldsymbol{M}}(z) \simeq \frac{1-\varphi}{z} + \frac{\varphi}{z} H(z) \tag{3.202}$$

where $H(z)$ *is the generating function defined as* $H(z) = 1 + \sum_{k=1}^{\infty} \frac{H_k}{z^k \varphi^k} = 1 + \frac{1}{\varphi} \sum_{k=1}^{\infty} \frac{\mathfrak{m}_k}{z^k}$.

Lemma 3.3.11. *Consider the matrix*

$$\begin{aligned} \boldsymbol{M} = \frac{1}{m} \mathbb{E}_{\boldsymbol{X}} \left[\phi(\boldsymbol{W}^{(\ell)} \boldsymbol{Y}^{(\ell)}) \phi(\boldsymbol{W}^{(\ell)} \boldsymbol{Y}^{(\ell)})^\top \right] \\ - \frac{1}{m^2 \sigma_x^2} \mathbb{E}_{\boldsymbol{x}} \left[\phi(\boldsymbol{W}^{(\ell)} \boldsymbol{Y}^{(\ell)}) \boldsymbol{X}^\top \right] \mathbb{E}_{\boldsymbol{x}} \left[\phi(\boldsymbol{W}^{(\ell)} \boldsymbol{Y}^{(\ell)}) \boldsymbol{X}^\top \right] \end{aligned} \tag{3.203}$$

and $\boldsymbol{Y}^{(\ell)} = \phi\left(\boldsymbol{H}^{(\ell-1)}\right)$ *as defined in Theorem 3.3.5, and consider a graph* G_3 *with* $2(p+q)$ *edges that is defined, via either an* i *or* μ *identification, by two graphs* G_1 *and* G_2 *with* $2p$ *and* $2q$ *edges. Then the asymptotic contribution,* $\mathfrak{m}_{p+q}^{G_3}$*, of the graph* G_3 *to the moment* $\mathfrak{m}_{p+q}$ *is equal to the product of the contributions of the block cycle graphs that define* G_3 *with a correction term. This is, they are joined with an* i*-identification*

$$\mathfrak{m}_{p+q}^{G_3} = \mathfrak{m}_p^{G_1} \frac{n_1}{n_1} \mathfrak{m}_p^{G_2} \tag{3.204}$$

and if they are joined with a μ*-identification*

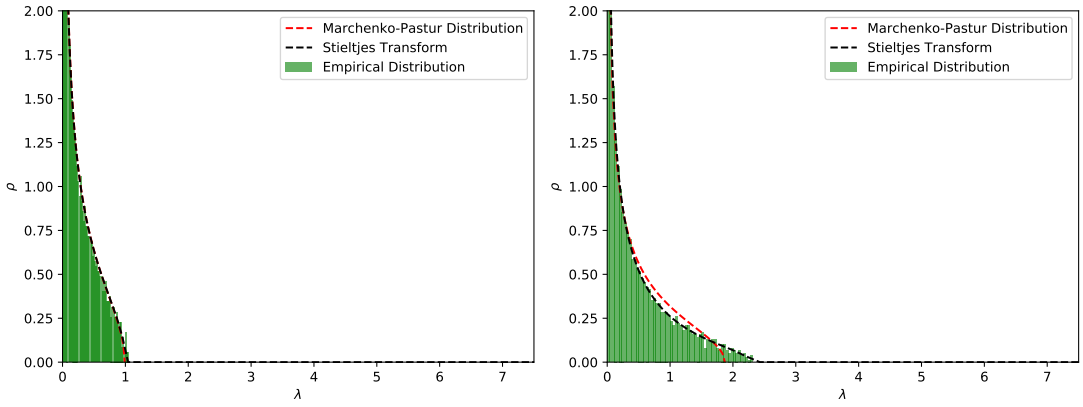
$$\mathfrak{m}_{p+q}^{G_3} = \mathfrak{m}_p^{G_1} \frac{n_1 \kappa_\alpha}{m} \mathfrak{m}_p^{G_2} \tag{3.205}$$

where if either one of the block cycles connected to the identification are of dimension 2*, then* $\alpha = b$ *or otherwise* $\alpha = c$ *and* $\kappa_b = 1 + \frac{\theta_1 \sigma_x^2 \sigma_w^2 \theta_3^2 + \sigma_x^2 \sigma_w^2 \theta_3^2 \theta_2 - \sigma_x^4 \sigma_w^4 \theta_3^4}{(\theta_1 - \sigma_x^2 \sigma_w^2 \theta_3^2)(\theta_2 - \sigma_x^2 \sigma_w^2 \theta_3^2)}$, $\kappa_c = 1 + \frac{2\sigma_x^2 \sigma_w^2 \theta_3^2 \theta_2 - \sigma_x^4 \sigma_w^4 \theta_3^4}{(\theta_2 - \sigma_x^2 \sigma_w^2 \theta_3^2)^2}$.

By using Lemmas 3.3.10 and 3.3.11 it is then possible to proof Theorem 3.3.5 by using the same exact process as for Theorem 3.3.1.

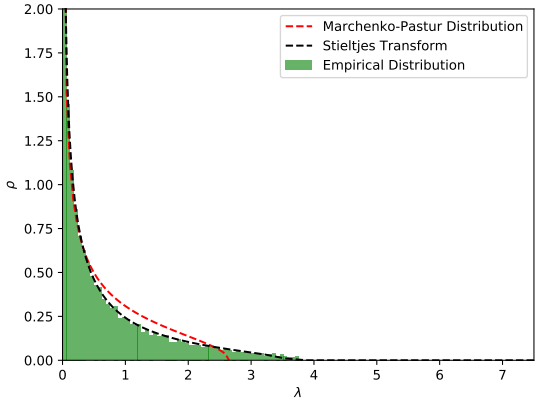
3.3.3 Numerical Results

In Figure 3.11, the empirical eigen-distribution of the matrices $\boldsymbol{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2} \boldsymbol{\Sigma}_{xh^{(\ell)}}^\top \boldsymbol{\Sigma}_{xh^{(\ell)}}$ at different layers are compared with the approximation that is retrieved by considering the Stieltjes transform in Theorem 3.3.5. Surprisingly, not considering the

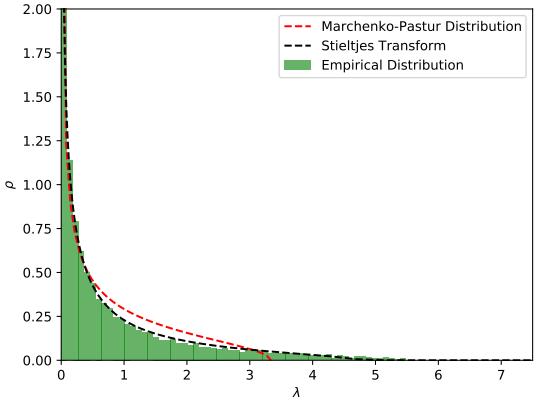


(a) 2nd Layer

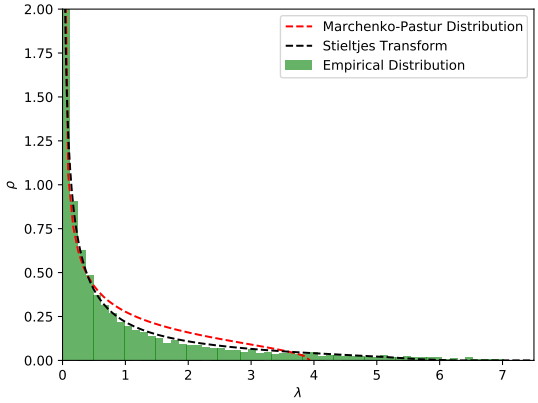
(b) 3rd Layer



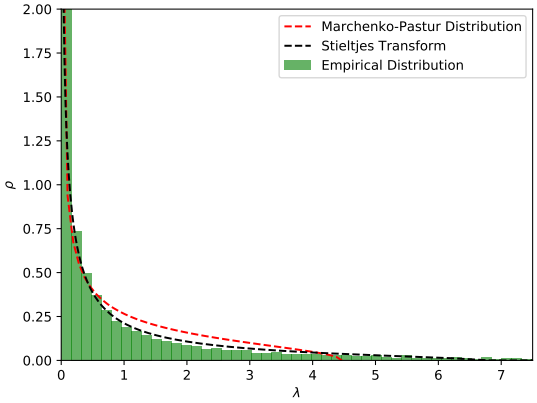
(c) 4th Layer



(d) 5th Layer



(e) 6th Layer



(f) 7th Layer

Figure 3.11: Empirical distribution of the eigenspectrum for the matrix $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^{\top}\mathbf{\Sigma}_{xh^{(\ell)}}$ with $n = 100$ and $(\sigma_w, \sigma_b, \sigma_n) = (2, 0, 0.001)$ at the first four layers.

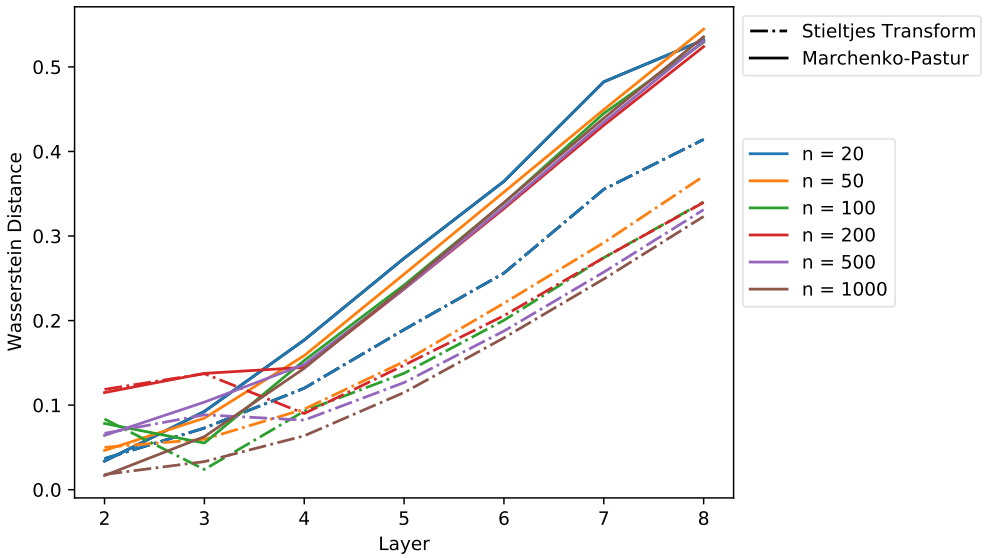


Figure 3.12: Comparison of the Wasserstein distance between the empirical distribution of $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$ for different network dimensions and the distribution attained with the Stieltjes transform (dash-dotted) and with the Marchenko-Pastur distribution (solid). For the second layer the Wasserstein distance decreases when n increases as expected. The modelling of deeper layers has a systematic error that is not accounted for and therefore the error is not decreasing with n. Further, the Marchenko-Pastur distribution does not converge to the empirical distribution with n for any layer.

way the signals correlate through the layers does not affect too much the approximation of the eigen-distribution of the matrix $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$. In Figure 3.9 we consider how the Wasserstein distance changes with the layers depth. As expected, the approximation relying on the Stieltjes transform performs better than the Marchenko-Pastur approximation besides for the second layer where the distribution is actually determined by the Marchenko-Pastur distribution. However, the approximation error does increase with the layers as the approximation of the hidden layers as having independent elements is naturally not accurate.

3.4 Improvement on the Approximation of the Mutual Information Lower Bound

In this section, the mutual information lower bound introduced in Chapter 2 is going to be computed with the eigen-distribution presented in the previous sections. However, we will first have to consider how the noise affects the distribution since the results in the previous sections were relative to $\sigma_n = 0$.

3.4.1 Noise Application

When considering the application of the noise, we need to remember that the expectation on which $\boldsymbol{\Lambda}_{h^{(\ell)}}$ and $\boldsymbol{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\boldsymbol{\Sigma}_{xh^{(\ell)}}^\top \boldsymbol{\Sigma}_{xh^{(\ell)}}$ are computed also considers the random vector $\mathbf{n}$. Therefore, as for the input $\mathbf{x}$ we need to consider an input matrix $\mathbf{N} = [\mathbf{n}_1, \dots, \mathbf{n}_m]$ rather than a vector and we will be considering the following cases

$$\Lambda_{\mathbf{h}^{(\ell+1)}} = \frac{\sigma_w^2}{n}\boldsymbol{W}\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})\phi(\mathbf{h}^{(\ell)})^\top]\boldsymbol{W}^\top + \mathbf{N}\mathbf{N}^\top \tag{3.206}$$

and

$$\begin{aligned}\boldsymbol{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\boldsymbol{\Sigma}_{xh^{(\ell)}}^\top \boldsymbol{\Sigma}_{xh^{(\ell)}} = \mathbf{W}^{(\ell)} \Big(&\mathbb{E}_{\mathbf{x}}[\phi(\mathbf{h}^{(\ell)})\phi(\mathbf{h}^{(\ell)})^\top] \\ &- \frac{1}{\sigma_x^2}\mathbb{E}_x[\phi(\mathbf{h}^{(\ell)})\mathbf{x}^\top]\mathbb{E}_x[\mathbf{x}\phi(\mathbf{h}^{(\ell)})^\top]\Big)\,\mathbf{W}^{(\ell)^\top} \\ &+ \mathbf{N}\mathbf{N}^\top.\end{aligned} \tag{3.207}$$

We are considering small noises, and for this reason we will assume that the noise at the previous layers do not affect the distribution significantly to be modelled.

The Stieltjes transform of the perturbation is

$$G_{\mathbf{N}\mathbf{N}^\top}(z) = \frac{1}{z - \sigma_n^2} \tag{3.208}$$

and therefore its inverse is

$$G^{-1}_{\mathbf{N}\mathbf{N}^\top}(z) = \frac{1}{z} + \sigma_n^2. \tag{3.209}$$

This allows to compute the R-transform of $\mathbf{N}\mathbf{N}^\top$

$$R_{\mathbf{N}\mathbf{N}^\top} = G^{-1}_{\mathbf{N}\mathbf{N}^\top}(z) - \frac{1}{z} = \sigma_n^2 \tag{3.210}$$

and thus by considering the implicit equation (3.37) we find that the Stieltjes transform of the perturbed matrix is

$$G_{\mathbf{A}+\mathbf{N}\mathbf{N}^\top}(z) = G_{\mathbf{A}}(z - \sigma_n^2). \tag{3.211}$$

This corresponds to a shift to the right of the distribution of the matrices that are used for the computation of the mutual information. Thanks to this shift the matrices $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$ are guaranteed of not being singular, and thus allow the computation of the mutual information.

Note that in the mean-field approximation of the mutual information lower bound, the noise was included in the computation of $q^{(\ell)}$, see Equation 2.51; there was no need on assuming that the noise was affecting only the last layer to simplify the computations.

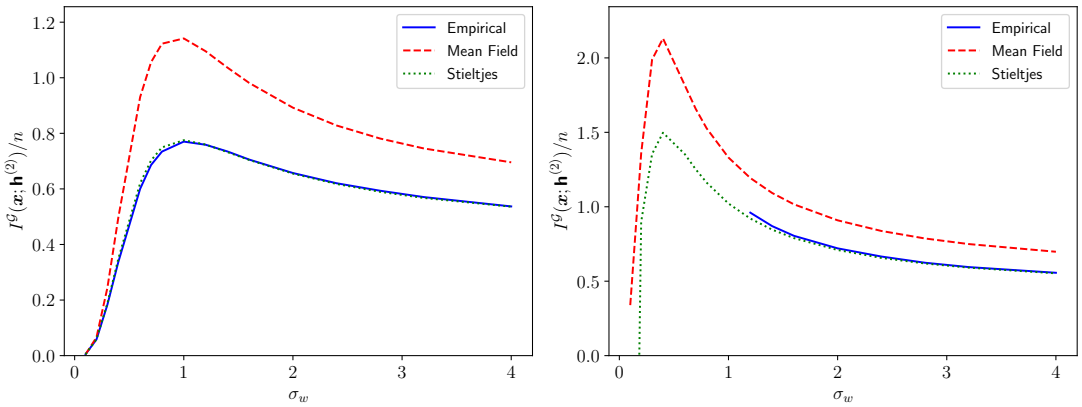
3.4.2 Estimation of the Mutual Information

Relying on the determination of the eigen-distribution of $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$ in Sections 3.2 and 3.3, and on the distribution sift caused by the noise application, it is now possible to numerically compute the lower bound of the mutual information.

In Figure 3.13 we compare the empirical mutual information lower bound for the second layer of a neural network to the approximation attained with the mean-field approximation and with the eigen-distribution approach when σ_w varies and $\sigma_b = 0$. The empirical lower bound generated by considering $1,000$ dimensional matrices with $1,000,000$ samples, is not always computable when σ_n is too small because the matrix $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$ is too ill conditioned and more samples are required. Nevertheless, for any computable combination of parameters, it is possible to notice how the analysis conducted in this Chapter allows to model exactly the behaviour, while the mean-field approximation overestimates the values.

The same comparison is illustrated in Figure 3.14 for the third layer. In this case the lower bound approximation generated with the eigen-distribution loses some accuracy as it was expected since we had to introduce some assumptions on the hidden layers when considering $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$. Interestingly, the distribution approximation agrees with the lower bound for large σ_w while the mean-field agrees with the lower bound for small σ_w.

Because of the computational complexity to compute the eigen-distribution of $\mathbf{\Lambda}_{h^{(\ell)}}$ for deeper layers, the analysis in this Chapter is limited to just the second and third layers.

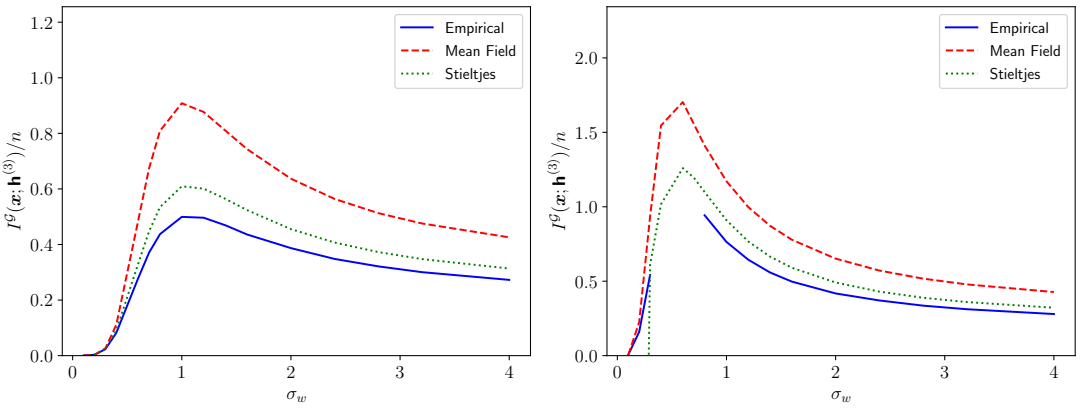


(a) $\sigma_n = 0.1$ (b) $\sigma_n = 0.01$

Figure 3.13: Comparison of the empirical mutual information lower bound between the input and the second hidden layer with its approximation with the mean-field approximation and with the Stieltjes transform. In (a) the results are relative to the noise $\sigma_n = 0.1$, while in (b) to the noise $\sigma_n = 0.01$. In (b) the empirical mutual information could not be computed for small σ_w since the matrices were too badly conditioned.

3.5 Considerations on the Eigen-distribution Approximation

The aim of this Chapter was to improve on the accuracy of the estimation of the lower bound of the mutual information by considering the eigen-distribution of $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$. We first introduced in Section 3.1 a simple approximation of the distributions with the Marchenko-Pastur distribution and discovered that the results in the mean-field approximation correspond to the values obtained when $\gamma \to 1$. We then considered improving the approximation by modelling the eigen-distribution of $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$ with the Stieltjes transform and the properties for free matrices. In the $\sigma_b = 0$ setting, in Section 3.2 we improved on [72] modelling the eigen-distribution of $\mathbf{\Lambda}_{h^{(\ell)}}$ exactly for deeper



(a) $\sigma_n = 0.1$ (b) $\sigma_n = 0.01$

Figure 3.14: Comparison of the empirical mutual information lower bound between the input and the third hidden layer with its approximation with the mean-field approximation and with the Stieltjes transform. In (a) the results are relative to the noise $\sigma_n = 0.1$, while in (b) to the noise $\sigma_n = 0.01$. In (b) the empirical mutual information could not be computed for small σ_w since the matrices were too badly conditioned.

layers, and in Section 3.3 we introduced an approximation for $\boldsymbol{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\boldsymbol{\Sigma}_{xh^{(\ell)}}^\top \boldsymbol{\Sigma}_{xh^{(\ell)}}$ which is especially accurate for the second layer. These results then allowed to improve the mutual information lower bound approximation as shown in Section 3.4.

Unfortunately, the implementation of the eigen-distribution based approximation is very time demanding which makes them unsuitable for numerially exploring the behaviour at deep layers. Despite this, the formulation of the eigen-spectra of $\boldsymbol{\Lambda}_{h^{(\ell)}}$ and $\boldsymbol{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\boldsymbol{\Sigma}_{xh^{(\ell)}}^\top \boldsymbol{\Sigma}_{xh^{(\ell)}}$ allows to identify the variables which govern the mutual information flow. Specifically, the flow is governed by

$$\theta_1^{(\ell)} = \int \frac{1}{\sqrt{2\pi}} \phi(\sigma_w \sqrt{q^{(\ell)}}\tilde{z})^2 e^{-\frac{\tilde{z}^2}{2}} d\tilde{z}, \tag{3.212}$$

$$\theta_2^{(\ell)} = \left(\int \frac{\sqrt{q^{(\ell)}}}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} \phi'(\sqrt{q^{(\ell)}}\tilde{z}) dz \right)^2, \tag{3.213}$$

$$\theta_3^{(\ell)} = \int \frac{\phi'\left(\sqrt{q^{(\ell)}}z\right)}{\sqrt{2\pi}} e^{-\frac{z^2}{2}} dz \tag{3.214}$$

and the rate at which $\left(\prod_{l=1}^{\ell} \theta_3^{(l)}\right)^2$ converges to 0 with ℓ defines the speed at which

the eigen-distribution of $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^{(\ell)}}^\top \mathbf{\Sigma}_{xh^{(\ell)}}$ converges to $\mathbf{\Lambda}_{h^{(\ell)}}$ and, therefore, the mutual information goes to 0.

Chapter 4

An Empirical Study of Derivative-Free-Optimisation Algorithms for Targeted Black-Box Attacks in Neural Networks[1]

After having presented in the previous Chapters our work on the propagation of information in random neural networks, we now focus on the work we conducted on the separate topic of adversarial examples.

Alongside the growing application of neural networks, there is a literature on their robustness which shows that it is often possible to subtly perturb the input image of a neural network in order to degrade its performance; these perturbed images are referred to as adversarial examples [32, 98]. For example, see [19, 25, 49, 94, 111] where road signals are perturbed so as to be wrongly interpreted by self driving cars that analyze images of them with neural networks. Methods to generate these adversarial examples are classified according to two main criteria [111]:

Adversarial Specificity establishes what the aim of the adversary is. In *non-targeted* attacks, the method perturbs the image in such a way that it is misclassified into any category other than the original one. While in *targeted*

[1]The material in the Chapter is taken from work published in the "Optimization and Engineering" Springer Journal [104]

settings, the adversary specifies a category into which an image should be misclassified.

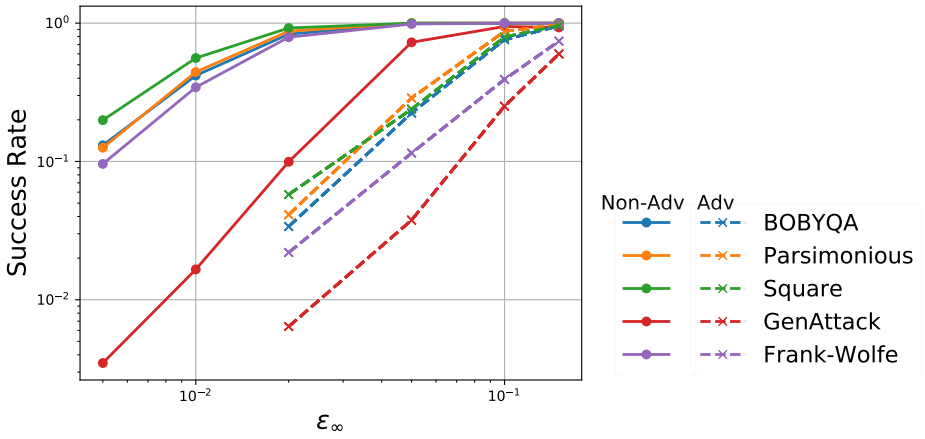
Adversary's Knowledge defines the amount of information available to the adversary. In *White-box* settings the adversary has complete knowledge of the network architecture and weights, while in the *Black-box* setting the adversary is only able to obtain the pre-classification output vector. The White-box setting allows for the use of gradients of a misclassification objective to efficiently compute the adversarial example [9, 14, 32], while the same optimization formulation of the Black-box setting requires use of a derivative free approach [4, 13, 41, 63].

In this work we consider the targeted black-box setting. In particular we follow [13] where:

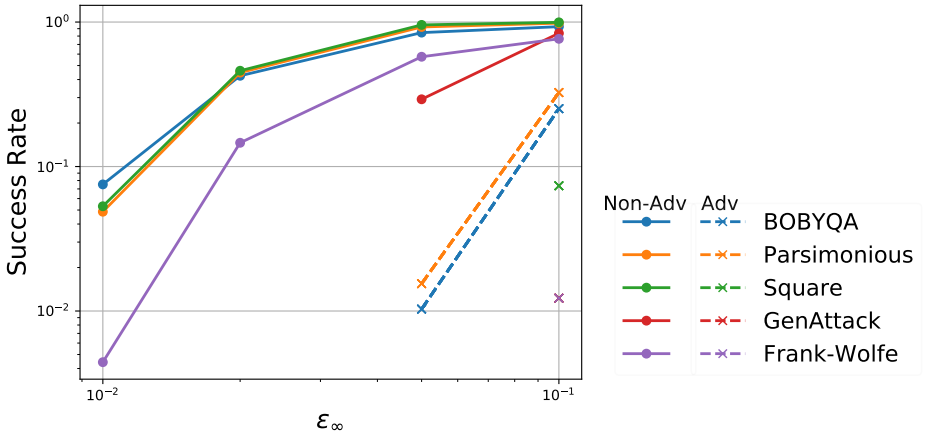
- the *perturbation*, which causes the network to change the classification, is bounded in magnitude by a specified ℓ^∞-norm, ε_∞, i.e. each pixel in the image cannot be perturbed by more than ${\varepsilon_\infty}^2$;
- the *number of queries* to the neural network needed to generate a targeted adversarial example should be as small as possible.

The Zeroth-Order-optimization (ZOO) algorithm proposed in [13] describes a Derivative Free optimization (DFO) method for computing adversarial examples in the black-box setting using a coordinate descent optimization method. At the time this was a substantial departure from previous black-box algorithms which trained a proxy neural network and then employ gradient based white-box attacks on the proxy network [70, 102]. It was demonstrated in [13] that these algorithms are especially effective when numerous adversarial examples are computed, but become less efficient when an individual adversarial example is considered. Following the introduction of ZOO, there have been numerous improvements using other model-free DFO based approaches, see for example [3–5, 12, 41, 42, 60]. Many of these

[2]In this work we always consider the pixels (for black and white images a pixel is the single channel input, while for RGB images we consider as different pixel the entries for each channel) to be normalised between 0 and 1. Therefore if we are allowed to perturb a pixel by ε_∞, then we are modifying its value by at most $(100\varepsilon_\infty)\%$.



(a) Cifar10



(b) ImageNet

Figure 4.1: The success rate of targeted attacks as a function of the perturbation's allowed ℓ^∞ magnitude for algorithms: GenAttack [4], Parsimonious [60], Square [5], Frank-Wolfe [12], and the BOBYQA based algorithm introduced here. Specifically for a ResNet50 network trained either on (a) the CIFAR10 or (b) the ImageNet dataset with (Adv) and without (Non-Adv) the defense by MadryLab [24]. An attack is considered successful if the method found the targeted adversarial example with less than 3,000 or 15,000 queries to the network trained on CIFAR and ImageNet dataset, respectively; Results for the case where the success rate is null, i.e. when no perturbations were successful, are excluded from the plot.

algorithms were developed in parallel, and so have not yet been bench-marked in a consistent setting, e.g. on the same network.

In this chapter, we present two frameworks for comparative evaluation of the existing algorithms that claim to have the fewest number of neural network queries to generate a successful attack. These are: GenAttack [4] which is based on a genetic direct-search method; Parsimonious algorithm [60], based on a combinatorial direct-search method on the vertices of the perturbation domain; the Square algorithm [5], based on a randomized direct-search method on the vertices of the perturbation domain; and the Frank-Wolfe algorithm [12] based on a momentum mechanism that approximates the gradient via finite differences. We also introduce a new algorithm built on a model-based DFO method [103]. In particular, we consider the *Bounded optimization BY Quadratic Approximation* (BOBYQA) [80] model-based DFO method which explicitly develops pseudo models to approximate the loss function in the optimization problem and then minimizes the loss function using methods from continuous optimization on the generated models. The aforementioned list of algorithms covers the leading classes of DFO algorithms for limited function evaluations, see e.g., [17, 50] for recent reviews of DFO methods. The two frameworks are structured as follows:

1. In the first setting we consider attacks on neural networks trained on CIFAR10 and ImageNet datasets, with or without the adversarial defense by MadryLab [24]; this is the canonical setup for the comparison of black-box attacks that was considered in previous literature. We illustrate in Figure 4.1 a measure of how the performance of the considered algorithms compare, while further refined measures of comparison are included in Section 4.3. We observe that the algorithms that limit the optimization domain to the ℓ^∞ perturbation boundary, i.e. the Parsimonious and Square algorithms, are consistently the most effective. In particular, the Square algorithm achieves the highest Success Ratio (SR) with a fixed maximum number of queries, except for when the neural networks have been adversarially trained, and the Parsimonious algorithm achieves the highest SR when a network is trained with the MadryLab defense. However, these results are relative to the current state-of-the-art defense in a field which is in continuous development

[20, 106] and newly proposed methods usually have a varying effect on the different attacking algorithms; for example the MadryLab defense [24] that we consider is most effective on Square algorithm in the ImageNet case.

2. In the second framework, the algorithms are allowed to perturb only a fraction of the pixels in the input; this is especially inspired by the structural defenses that transform the input in the wavelet space [36]. This framework allows us to understand the sensitivity of different algorithms to choices such as initialization, experimental protocol, dataset, and adversarial training. Our results demonstrate that the Parsimonious, Square, and BOBYQA based algorithms alternatively perform the best for different maximum perturbation energies.

The results in this chapter show that the most likely algorithm to find an adversarial example varies according to the considered setting; the type of dataset, the defense, and the perturbation energy bound have a varying impact on the different algorithms. As a consequence of these experiments, new algorithms should be compared to the state-of-the-art in a variety of settings as done here, and the effectiveness of an adversarial defense should be tested with a variety of algorithms, including the BOBYQA based algorithm that we first introduced in [103].

The outline of the chapter is as follows: in Section 4.1 we present how an adversarial example is generated by solving an optimization problem, and how DFO methods fit in this context. We also introduce the model-based BOBYQA algorithm. In Section 4.2 we present two popular techniques used in existing methods to improve the efficiency and scalability to high dimensional inputs. Section 4.3 presents the experimental setup and a comparative analysis of existing algorithms along with a focus on our proposed BOBYQA based algorithm. We close with some concluding remarks in Section 4.4.

4.1 Adversarial Examples Formulated as an optimization Problem

In classification tasks, a neural network outputs a vector whose length is equal to the number of classes and the neural network parameters are trained to match the

maximum element of the given output to the correct class of the input. Adversarial perturbations are obtained by modifying the input in such a way that the maximum element of neural network output corresponds to a target class different from the original one.

Consider a neural network $F : \mathcal{X} \to \mathcal{Y}$ from input space $\mathcal{X}$ to output space $\mathcal{Y}$ of classes. A targeted adversarial perturbation $\boldsymbol{\eta}$ to an input $\mathbf{X} \in \mathcal{X}$ has the property that it changes the classification to a specified target class t, i.e $F(\mathbf{X}) = c$ and $F(\mathbf{X} + \boldsymbol{\eta}) = t \neq c$.

Following the formulation in [4]; given an input space $\mathcal{X} = [l, u]^n$, with l and u being respectively the minimum and maximum values of the interval in which the pixels may vary, an output space $\mathcal{Y} = \{1, \ldots, n_c\}$, where n_c is the number of classes, a maximum energy budget ε_∞, and a suitable loss function $\mathcal{L}$, then the task of computing the adversarial perturbation $\boldsymbol{\eta}$ can be cast as an optimization problem such as

$$\begin{aligned} &\min_{\boldsymbol{\eta}} \mathcal{L}(\mathbf{X}, \boldsymbol{\eta}) \\ &\text{s.t. } \|\boldsymbol{\eta}\|_\infty \leq \varepsilon_\infty; \\ &\quad [\mathbf{X} + \boldsymbol{\eta}]_j \geq l \qquad \forall j \in 1, ..., n \\ &\quad [\mathbf{X} + \boldsymbol{\eta}]_j \leq u \qquad \forall j \in 1, ..., n \end{aligned} \tag{4.1}$$

where the final two inequality constraints are due to the perturbed image being still an image, i.e. $(\mathbf{X} + \boldsymbol{\eta}) \in \mathcal{X}$. Denoting the pre-classification output vector by $f(\mathbf{X})$, i.e. $F(\mathbf{X}) = \arg\max f(\mathbf{X})$, then the misclassification of $\mathbf{X}$ to target label t is achieved by $\boldsymbol{\eta}$ if $f(\mathbf{X} + \boldsymbol{\eta})_t \geq \max_{j \neq t} f(\mathbf{X} + \boldsymbol{\eta})_j$. As demonstrated in [4, 9, 13], in this study we consider the following loss function for computing $\boldsymbol{\eta}$ in (4.1)

$$\mathcal{L}(\mathbf{X}, \boldsymbol{\eta}) = \log\left(\Sigma_{j \neq t} f(\mathbf{X} + \boldsymbol{\eta})_j\right) - \log\left(f(\mathbf{X} + \boldsymbol{\eta})_t\right). \tag{4.2}$$

Not having access to the internal parameters of the neural network, the gradient of the loss over the input space cannot be readily computed and instead the adversarial perturbation is found using specially adapted DFO algorithms.

4.1.1 Derivative Free Optimization for Adversarial Examples

Derivative free optimization is a well developed field with numerous classes of methods, see [17] and [50] for reviews on DFO principles and algorithms. Example classes of such methods include: direct search methods such as simplex, model-based methods, hybrid methods such as finite differences or implicit filtering, as well as randomized variants of the aforementioned and methods specific to convex or noisy objectives. For the generation of adversarial examples, the algorithms that we consider rely on three types of DFO methods:

- those where the gradient is computed via finite differences, either by sampling all the canonical directions as in ZOO attack [13] or random directions as in the Frank-Wolfe algorithm [12];
- those where the solution is thought to be in one of the vertices of the ℓ^∞ domain, i.e. $\boldsymbol{\eta}_i \in \{-\varepsilon_\infty, \varepsilon_\infty\}$ for any i. The Parsimonious algorithm [60] implements a combinatorial direct-search within the different possible vertices, initializing the perturbation to $-\varepsilon_\infty$ for all the pixels and then switching collections of them to $+\varepsilon_\infty$, when such an action decreases the loss function. The Square algorithm [5] instead implements a randomized direct-search method where square blocks of pixels are iteratively perturbed to be either $+\varepsilon_\infty$ or $-\varepsilon_\infty$;
- those where a direct search over the perturbation domain is performed using a genetic method such as GenAttack [4].

The optimization formulation in (4.1) is amenable to virtually all DFO methods, making it unclear which of the methods would be most effective in this context. Further, model-based methods are notably missing from the aforementioned list. Thus for completeness, we introduce an algorithm relying on a model-based method; specifically, BOBYQA is considered given its proven effectiveness in solving complex problems such as climate modelling [101].

4.1.2 Model-Based DFO

Given a set of q samples $\mathcal{Y} = \{\mathbf{y}^1, ..., \mathbf{y}^q\}$ with $\mathbf{y}^i \in \mathbb{R}^n$, model-based DFO methods start by identifying the minimizer of the objective among the samples at iteration k, $\mathbf{x}^k = \arg\min_{\mathbf{y}\in\mathcal{Y}} \mathcal{L}(\mathbf{y})$. Following this, a model for the objective function $\mathcal{L}$ is constructed, typically centered around the minimizer. In its simplest form one uses a polynomial approximation to the objective, such as a quadratic model centered in $\mathbf{x}^k$

$$m_k(\mathbf{x}^k + \mathbf{p}) = a_k + \mathbf{c}_k^\top \mathbf{p} + \frac{1}{2}\mathbf{p}^\top \mathbf{M}_k \mathbf{p}, \tag{4.3}$$

with $a_k \in \mathbb{R}$, $\mathbf{c}_k$, $\mathbf{p} \in \mathbb{R}^n$, and $\mathbf{M}_k \in \mathbb{R}^{n\times n}$ being also symmetric. In a white-box setting one would set $\mathbf{c}_k = \nabla\mathcal{L}(\mathbf{x}^k)$ and $\mathbf{M}_k = \nabla^2\mathcal{L}(\mathbf{x}^k)$, but this is not feasible in the black-box setting as we do not have access to the derivatives of the objective function. Thus at each iteration k, the parameters a_k, $\mathbf{c}_k$ and $\mathbf{M}_k$ are usually defined by imposing interpolation conditions

$$m_k(\mathbf{y}^i) = \mathcal{L}(\mathbf{y}^i) \quad \forall i \in 1, 2, \ldots, q, \tag{4.4}$$

and when $n + 1 \leq q < 1 + n + n(n + 1)/2$ (i.e. the system of equations is underdetermined) the model could be set as linear by imposing $\mathbf{M}_k = \mathbf{0}$ for any k [65], or the interpolation conditions could be considered as the constraint in an optimisation problem, as done in the BOBYQA method by [80] presented in the following subsection. The objective model (4.3) is considered to be a good estimate of the objective in a neighborhood referred to as a trust region. Once the model m_k is generated, the update step $\mathbf{p}$ is computed by solving the trust region problem

$$\begin{aligned} \min_{\mathbf{p}} \quad & m_k(\mathbf{x}_k + \mathbf{p}) \\ \text{s.t.} \quad & \|\mathbf{p}\| \leq \Delta, \end{aligned} \tag{4.5}$$

where Δ is the radius of the region where we believe the model to be accurate, for more details on trust region methods see [65]. The new point $\mathbf{x}_k + \mathbf{p}$ is added to $\mathcal{Y}$ and a prior point is potentially removed to improve the accuracy of the model according to geometric considerations, such as the poisedness of the sample set which minimizes the potential for degeneracy of the model, [86] for details. In this chapter, we consider an exemplary model-based method called BOBYQA.

4.1.2.1 BOBYQA

The *Bound Optimization BY Quadratic Approximation* (BOBYQA) method, introduced in [80], updates the parameters of the model $a, \mathbf{c}$, and $\mathbf{M}$, in each iteration in such a way as to minimize the change in the quadratic term $\mathbf{M}_k$ between iterates while otherwise fitting the sample values:

$$\begin{aligned} &\min_{a_k, \mathbf{c}_k, \mathbf{M}_k} \|\mathbf{M}_k - \mathbf{M}_{k-1}\|_F^2 \\ &\text{s.t.} \quad m_k(\mathbf{y}^i) = \mathcal{L}(\mathbf{y}^i), \qquad \forall i \in 1, 2, \ldots, q, \end{aligned} \tag{4.6}$$

with $n+1 < q < 1+n+n(n+1)/2$ and $\mathbf{M}_k$ initialized as the zero matrix. When the number of parameters $q = n+1$ then the model is considered as linear with $\mathbf{M}_k$ set as zero. At each iteration the set $\mathcal{Y}$ is updated with the insertion of the new point, $\mathbf{x}_k + \mathbf{p}$, and the removal of the sample which affects the most negatively the model accuracy. The two main notions used to determine which sample to remove are distance from the new sample and a measure to minimize the potential for degeneracy of the model, considerations that are built on concepts of stability [80] and of poisedness [86], for more details refer to [82, Chapter 6]. This process of maintaining a fixed number of samples insures that the dimension of $\mathcal{Y}$ is fixed.

4.2 Improving Efficiency and Computational Scalability

Because of the high number of pixels in the input images[3], the generation of adversarial examples involves solving a high dimensional problem, which makes the use of any DFO method impractical; for instance, the application of the BOBYQA method requires the solution of (4.6) which scales in memory allocation at least quadratically with the input dimension, and thus is computationally too expensive. Consequently, the implementation of DFO based adversarial algorithms relies on strategies to reduce the dimensionality of the problem, this improves the computational scalability along with the efficiency, as demonstrated experimentally. Instead of solving (4.1) for $\boldsymbol{\eta} \in \mathbb{R}^n$ directly, the DFO based algorithms consider

[3]For example, the inputs in the ImageNet dataset have 268,203 pixels.

variations of the domain sub-sampling and/or hierarchical liftings techniques. Domain sub-sampling iteratively sweeps over batches of $b \ll n$ variables, while hierarchical lifting clusters and perturbs variables simultaneously, as described in following sections.

4.2.1 Domain Sub-Sampling

The simplest version of domain sub-sampling consists of partitioning the input dimension into smaller disjoint domains and optimizing the loss function in each of them sequentially. That is, in an n dimensional problem, one considers $k = \lceil n/b \rceil$ sets of integers, $\{\Omega^j\}_{j=1}^k$, of size $b \ll n$ which are disjoint and which cover all of $[n]$. Then (4.1) is solved sequentially on the dimensions identified by the sets Ω^j. This is possible since the optimization domain is box like, i.e. $\boldsymbol{\eta} \in [l, u]^n$, and each dimension's bound is independent from the others. Formally, rather than solving (4.1) for $\boldsymbol{\eta} \in \mathbb{R}^n$ directly, for each of $j = 1, \ldots, k$ one *sequentially* solves for the $\boldsymbol{\eta}^j \in \mathbb{R}^n$ variables which are only non-zero for entries in Ω^j. The resulting sub-domain perturbations $\boldsymbol{\eta}^j$ are then summed to generate the full perturbation $\boldsymbol{\eta} = \sum_{j=1}^k \boldsymbol{\eta}^j$, see Figure 4.2 as an example. That is, the optimization problem (4.1) is adapted to repeatedly looping over $j = 1, \ldots, k$:

$$
\begin{aligned}
\min_{\boldsymbol{\eta}^j} \quad & \mathcal{L}\left(\mathbf{X} + \sum_{h \neq j} \boldsymbol{\eta}^\ell, \boldsymbol{\eta}^j\right) \\
\text{s.t.} \quad & \left\| \sum_{h=1}^k \boldsymbol{\eta}^h \right\|_\infty \leq \varepsilon_\infty; \\
& \left[\mathbf{X} + \sum_{h=1}^k \boldsymbol{\eta}^h\right]_r \geq l \qquad \forall r \in \Omega^j; \\
& \left[\mathbf{X} + \sum_{h=1}^k \boldsymbol{\eta}^h\right]_r \leq u \qquad \forall r \in \Omega^j,
\end{aligned}
\tag{4.7}
$$

where the sets $\{\Omega^j\}_{j=1}^k$ are usually computed again once j is equal to k, and the sub-domain perturbations $\boldsymbol{\eta}^j$ are initialized as null.

We identified three possible ways of selecting the sub-domains $\{\Omega^j\}_{j=1}^k$;

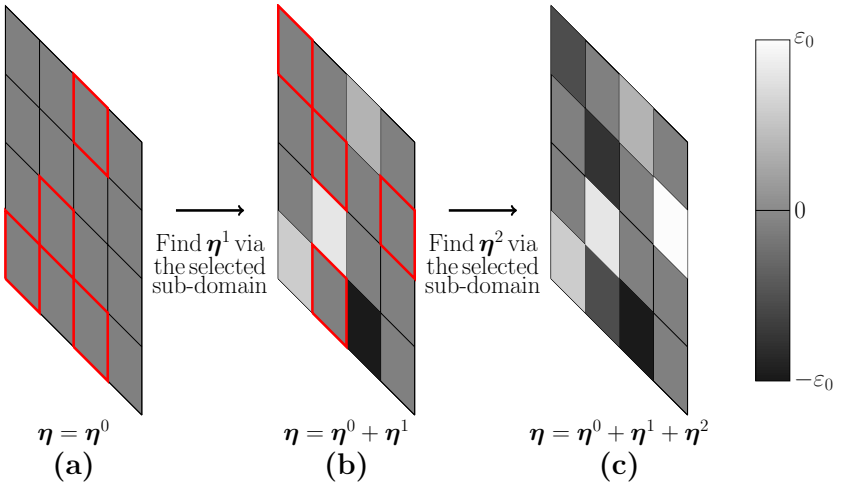


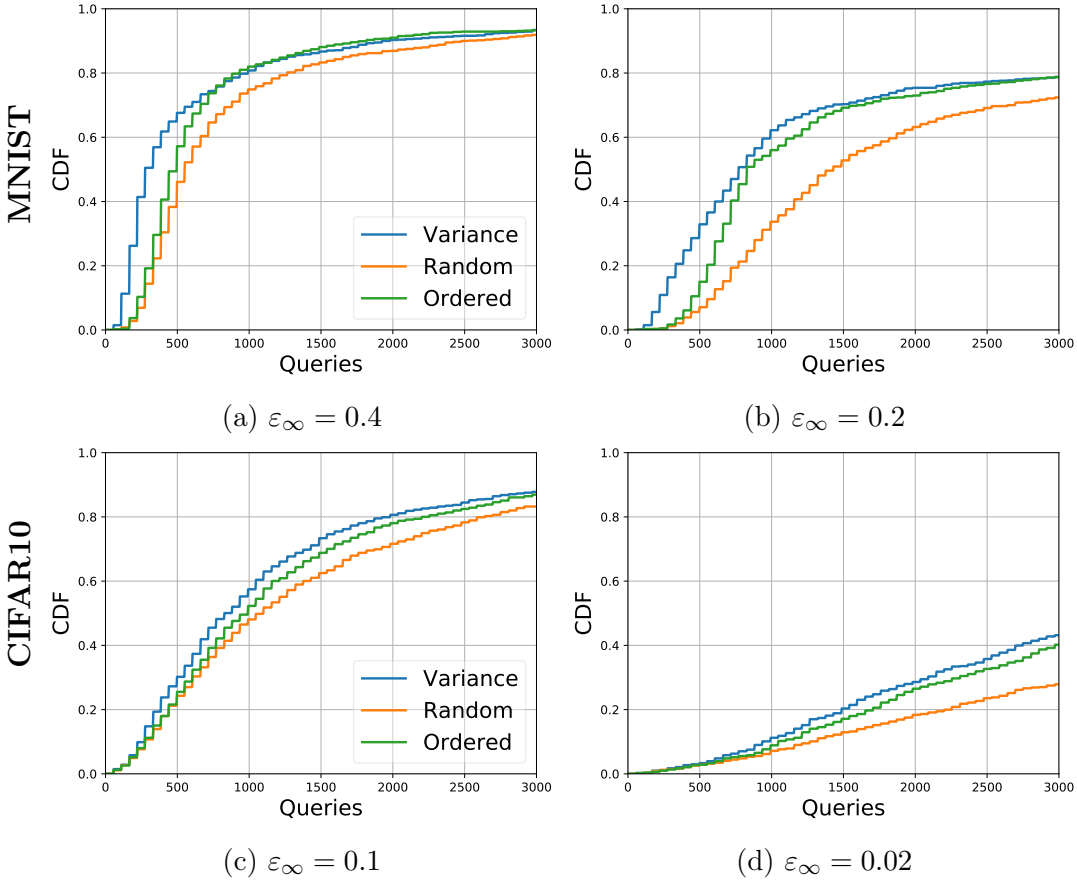
Figure 4.2: Example of how the perturbation $\boldsymbol{\eta}$ evolves through the iterations when an image in $\mathbb{R}^{4\times4}$ is attacked. In (a) the perturbation is $\boldsymbol{\eta} = \boldsymbol{\eta}^0$ and a sub-domain of $b = 4$ pixels (in red) is selected. Once the optimal perturbation $\boldsymbol{\eta}^1$ in the selected sub-domain is found, the perturbation is updated in (b) and a new sub-domain of dimension b is selected. The same is repeated in (c).

- In *Random Sampling* one considers at each iteration a different random sub-samplings of the domain, i.e. $k = 1$. The ZOO algorithm used this kind of sampling [13].

- In *Ordered Sampling* one generates a random disjoint partitioning of the domain, i.e. $k = \lceil n/b \rceil$ and $\Omega_j \cap \Omega_l = \emptyset$ for any j and l. A new partitioning is generated when each variable has been optimized over once. This sampling is implemented in the Parsimonious algorithm.

- In *Variance Sampling* one still generates a random disjoint partitioning of the domain, but chooses the sub-samplings sets $\{\Omega^j\}_{j=1}^k$ in order to optimize over the dimensions that have highest local variance in intensity first. Specifically, the variables are ordered by the variance in intensity among the 8 neighboring variables (e.g. pixels) in the same color channel of the input $\mathbf{X}$. The sets $\{\Omega^j\}_{j=1}^k$ are further reinitialized after each loop through $j = 1, \ldots, k$.

The sub-sampling of the domain affects the efficiency with which an algorithm successfully finds an adversarial example. For instance, in Figure 4.3 we compare how these different sub-sampling techniques affect the BOBYQA based algorithm



(a) $\varepsilon_\infty = 0.4$ (b) $\varepsilon_\infty = 0.2$

(c) $\varepsilon_\infty = 0.1$ (d) $\varepsilon_\infty = 0.02$

Figure 4.3: Cumulative distribution function of successfully perturbed images as a function of number of queries by the BOBYQA based algorithm attacking neural networks trained on the MNIST and the CIFAR10 datasets. In each image the effectiveness of different sub-sampling methods in generating a successful adversarial example is shown for different values of maximum perturbation energies ε_∞. See [103] for details about experimental setup.

when generating adversarial example for the MNIST and CIFAR10 dataset. It can be observed that variance sampling consistently has a higher success rate cumulative distribution function as compared with random and ordered sampling. This suggest that pixels belonging to high-contrast regions are more influential than the ones in low-contrast ones, and hence variance sampling is the preferable ordering.

To simplify the notation in the following section, the optimization variable is considered to be $\boldsymbol{\eta}^j = \boldsymbol{\Omega}^j \tilde{\boldsymbol{\eta}}^j$ where $\tilde{\boldsymbol{\eta}}^j \in \mathbb{R}^b$ and $\boldsymbol{\Omega}^j \in \mathbb{R}^{n \times b}$ is such that $[\boldsymbol{\Omega}^j]_{pq}$ is

Algorithm 3: GENERATE_SAMPLING_MATRIX($\hat{\mathbf{X}}$,n_ℓ,b,j)

$\mathbf{\Omega} \leftarrow \mathbf{0} \in \mathbb{R}^{n_\ell \times b}$, $\boldsymbol{\mu} \leftarrow \mathbf{0} \in \mathbb{R}^{n_\ell}$,$\boldsymbol{v} \leftarrow \mathbf{0} \in \mathbb{R}^{n_\ell}$.
for $i = 1, \ldots, b$ **do**
    $\boldsymbol{\mu}(i) = \sum_{j \in \text{Neighbours}(i)} \hat{\mathbf{X}}(j)$.
    $\mathbf{v}(i) = \sum_{j \in \text{Neighbours}(i)} \hat{\mathbf{X}}(j)^2 - \boldsymbol{\mu}^2(i)$. *# computation of the variance around the variable.*
end for
$\mathbf{s} \leftarrow$ argsort $\mathbf{v}$.
for $i = 1, \ldots, b$ **do**
    $\mathbf{\Omega}(\mathbf{s}[i + j \times b], [i]) = 1$.
end for
Return $\mathbf{\Omega}$.

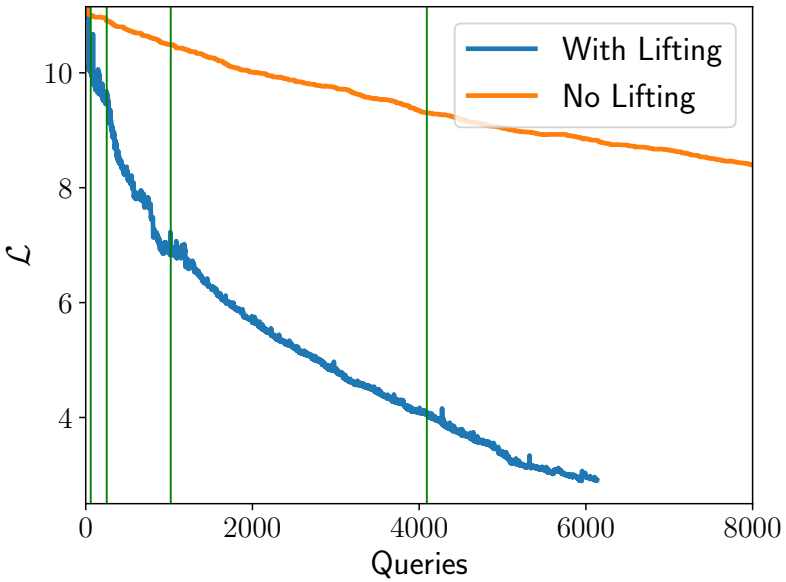


Figure 4.4: Impact of hierarchical lifting approach on Loss function (4.2) as a function of the number of queries to a ResNet50 trained on ImageNet dataset to find the adversarial example for a single image with the BOBYQA based method. The green vertical lines correspond to changes of hierarchical level, which entail an increase in the dimension of the optimization space.

one if the qth element of Ω^j is p, zero otherwise. The implementation of variance sampling method at iteration j in a domain of dimension n_ℓ is summarized in Algorithm 1.

4.2.2 Hierarchical Lifting

Authors of ZOO attack [13] demonstrated that fewer queries are required to find adversarial example when pixels are considered in clusters, and not independently. This lead to the hierarchical lifting approach where one optimizes over increasingly higher dimensional spaces at each step, referred here as level ℓ; Figure 4.4 shows how effective this approach is when implementing the BOBYQA based algorithm. These low dimensional spaces are lifted to the image space via a linear lifting, where at each level ℓ a linear lifting $\mathbf{D}^\ell : \mathbb{R}^{n_\ell} \to \mathbb{R}^n$ is considered and a perturbation $\hat{\boldsymbol{\eta}}_\ell \in \mathbb{R}^{n_\ell}$ is found to be added to the full perturbation $\boldsymbol{\eta}$, according to

$$\boldsymbol{\eta} = \sum_{j=0}^{\ell} \boldsymbol{\eta}_j = \sum_{j=0}^{\ell} \mathbf{D}^j \hat{\boldsymbol{\eta}}_j. \tag{4.8}$$

Here $\boldsymbol{\eta}_0$ is initialized as $\underline{\mathbf{0}}$ and the perturbations $\boldsymbol{\eta}_j$ of the previous layers are considered as fixed. An example of how this works is illustrated in Figure 4.5. The hierarchical lifting considered here is analogous to the derivative-based Recursive Multiscale Trust-Region method in [34]. Our piece-wise constant lifting is substantially simpler than [34] in that we only progress from coarse to fine grids as opposed to "W" and "V" cycles between scales; this simplicity is beneficial for the misclassification application here where our aim is to minimize the number of function queries used by the DFO method.

All the methods considered in this work rely on ideas which can be interpreted through this approach. The algorithms that we consider in this work rely on two kinds of linear lifting $\mathbf{D}^\ell$ differentiated by the way each scalar in $\hat{\boldsymbol{\eta}}$ is associated to a set of pixels in the original image domain $\mathbb{R}^n$; namely the random and the block liftings. The former relates a random set of pixels of the original image to each hyper-variable; this forces the perturbation to be of high-frequency nature, as illustrated in Figure 4.6(a), which several articles indicate as being the most effective [33, 35, 90]. The GenAttack and Frank-Wolfe algorithms use a variation of this kind of lifting. The latter instead is based on interpolation operations; a sorting matrix $\mathbf{S}^\ell : \mathbb{R}^{n_\ell} \to \mathbb{R}^n$ is applied such that every index of $\hat{\boldsymbol{\eta}}_\ell$ is uniquely associated to a node of a coarse grid masked over the original image. Afterwards, an interpolation $\mathbf{L}^\ell : \mathbb{R}^n \to \mathbb{R}^n$ is implemented over the values in the coarse grid,

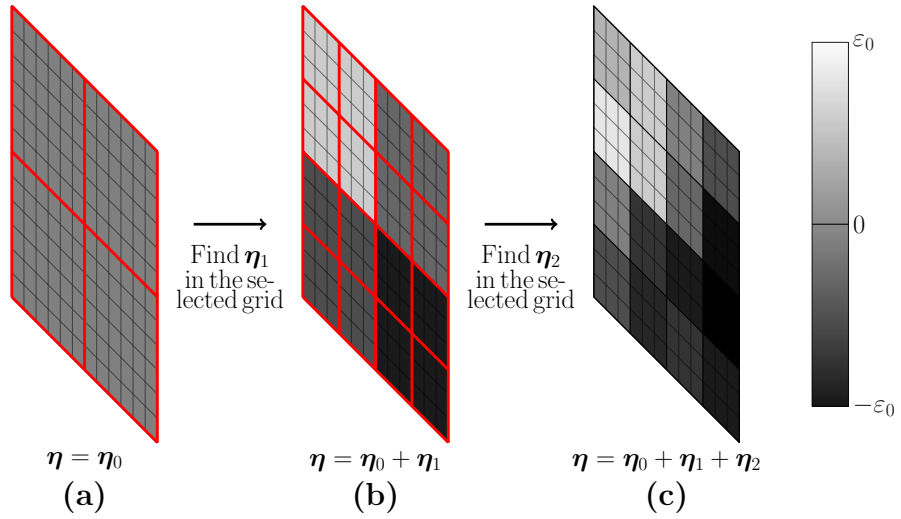
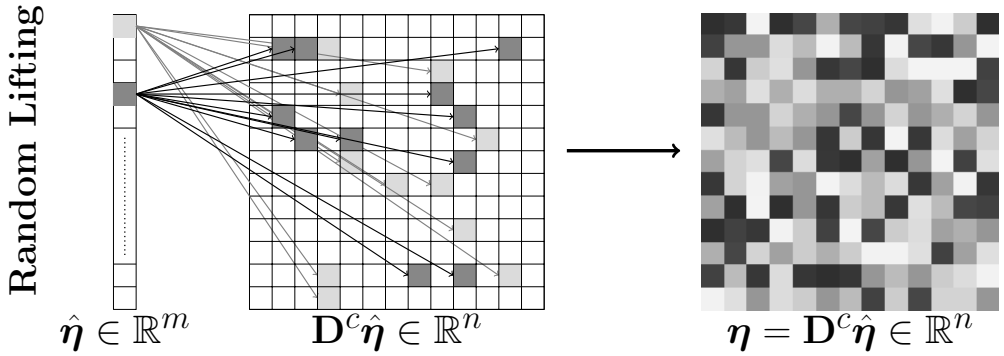


Figure 4.5: Example of how the perturbation $\boldsymbol{\eta}$ is generated in a hierarchical lifting method with $n_1 = 4$ and $n_2 = 16$ on an image in $\mathbb{R}^{12\times 12}$. In (a) the perturbation is $\boldsymbol{\eta} = \boldsymbol{\eta}_0$ and the boxes generated via the grid of dimension n_1 are highlighted in red. Once the optimal perturbation $\boldsymbol{\eta}_1$ is found, the perturbation is updated in (b) and the image is further divided with a grid with n_2 blocks. The final solution obtained after optimization is shown in (c).

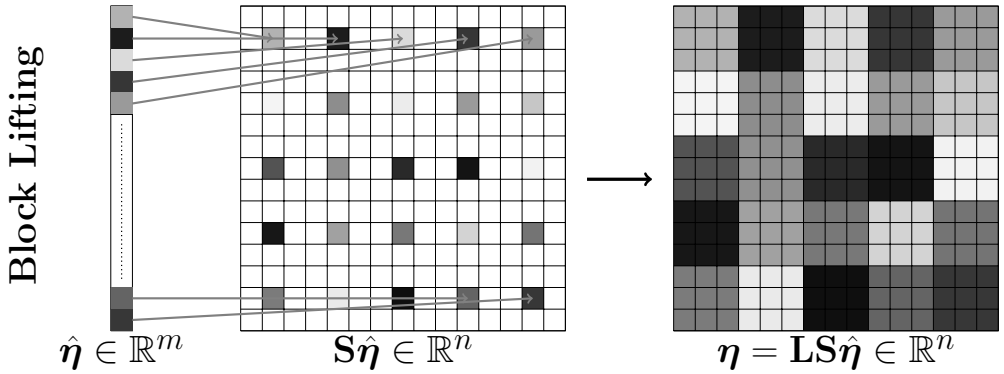
i.e. $\boldsymbol{\eta}_\ell = \mathbf{L}^\ell \mathbf{S}^\ell \hat{\boldsymbol{\eta}}_\ell = \mathbf{D}^\ell \hat{\boldsymbol{\eta}}_\ell$. Both Square and Parsimonious algorithms implement hierarchical lifting with the piece-wise constant interpolation, here referred to as block lifting. At the lower levels the interpolation lifting generates low frequency perturbations, as illustrated in Figure 4.6(b).

Since n_ℓ may still be very high, for each level ℓ domain sub-sampling is also applied considering $\hat{\boldsymbol{\eta}}_\ell = \sum_{j=0}^{k} \tilde{\boldsymbol{\eta}}_\ell^j$. In the piece-wise constant case with variance sampling, the blocks are ordered according to the variance of mean intensity among neighboring blocks, in contrast to the variance within each block as suggested in [13]. Consequently, at each level the adversarial example is found by solving the following iterative problem

$$\begin{aligned} \min_{\tilde{\boldsymbol{\eta}}_\ell^j} \quad & \mathcal{L}\left(\mathbf{X} + \bar{\boldsymbol{\eta}}, \mathbf{D}^\ell \boldsymbol{\Omega}^k \tilde{\boldsymbol{\eta}}_\ell^j\right) \\ \text{s.t.} \quad & \left\| \bar{\boldsymbol{\eta}} + \mathbf{D}^\ell \boldsymbol{\Omega}^k \tilde{\boldsymbol{\eta}}_\ell^j \right\|_\infty \leq \varepsilon_\infty \\ & \left[\mathbf{X} + \bar{\boldsymbol{\eta}} + \mathbf{D}^\ell \boldsymbol{\Omega}^k \tilde{\boldsymbol{\eta}}_\ell^j\right]_r \geq l \qquad \forall r \in \{1, ..., n\} \\ & \left[\mathbf{X} + \bar{\boldsymbol{\eta}} + \mathbf{D}^\ell \boldsymbol{\Omega}^k \tilde{\boldsymbol{\eta}}_\ell^j\right]_r \leq u \qquad \forall r \in \{1, ..., n\}, \end{aligned} \tag{4.9}$$



(a)



(b)

Figure 4.6: Examples for (a) random and (b) block liftings. In the random case each pixel in the perturbation is associated to just one element of $\hat{\boldsymbol{\eta}}_\ell$. Block lifting uses a piece-wise constant interpolation $\mathbf{L}$ over a coarse grid $\mathbf{S}\hat{\boldsymbol{\eta}}_\ell$ and each block is associated uniquely to one of the variables in $\hat{\boldsymbol{\eta}}_\ell$. In both cases, the lifting $\mathbf{D}$ is such that each element $\mathbf{D}_{ij}$ is either 1 or 0.

Algorithm 4: GENERATE_LIFTING(n_ℓ,n)

1: $\mathbf{D} \leftarrow \mathbf{0} \in \mathbb{R}^{n \times n_\ell}$
2: **for** $i = 1, \ldots, n_\ell$ **do**
3: Generate set of pixels S that are in the block associated to the i-th element of the n_ℓ dimensional super-grid.
4: **for** $j \in S$ **do**
5: $\mathbf{D}(i, j) = 1$.
6: **end for**
7: **end for**
8: Return $\mathbf{D}$.

where $\bar{\boldsymbol{\eta}} = \sum_{i=0}^{\ell-1} \boldsymbol{\eta}_i + \mathbf{D}^\ell \sum_{m \neq j} \hat{\boldsymbol{\eta}}_\ell^m$. Algorithm 4 gives an implementation of the block lifting matrix when in the grid has dimension n_ℓ.

4.3 Comparison of Derivative Free Methods

In this section, we compare algorithms based on a selection of state-of-the-art DFO methods. In particular we consider BOBYQA based algorithm [103], GenAttack algorithm [4], Parsimonious algorithm [60], Square algorithm [5] and Frank-Wolfe algorithm [12] in the following two frameworks:

- Section 4.3.3 considers the canonical setup for black-box adversarial attacks on which the considered algorithms have been tuned in their respective chapters. Specifically, we consider attacks on networks trained adversarially or not on CIFAR10 and ImageNet, two popular datasets in the literature, and with no further defense implemented.

- Section 4.3.4 considers a setup that simulates structural defenses on which the different algorithms were not tuned. We limit the perturbation to a fixed number of pixels with high variance in intensity considering attacks on a network non-adversarially trained on the CIFAR10 dataset.

The performance of all algorithms is measured in terms of the distribution of queries needed to successfully find adversaries to identical networks given a fixed ℓ^∞ perturbation constraint and the same input images. In particular, the algorithms are compared according to the cumulative fraction of images successfully

misclassified (abridged by CDF for cumulative distribution function) as a function of the number of queries to the neural network, which corresponds to the data profile comparison measure introduced in [61]. For each experimental setting $\mathcal{A}$, the single attacks are denoted by a, and the following variable is introduced

$$t_a = \#\text{ of queries to find an adversarial attack} \tag{4.10}$$

that is set to infinity in case the adversarial example is not found. Thus, the CDF for a number of queries α is

$$CDF(\alpha) = \frac{1}{|\mathcal{A}|} size\{a \in \mathcal{A} : t_a \leq \alpha\}. \tag{4.11}$$

4.3.1 Parameter Setup for Algorithms

The experiments use publicly available implementations for the GenAttack [4], Parsimonious [60], Square [5], and Frank-Wolfe [12] algorithms[4] using the same hyper-parameter setting and hierarchical lifting approach as suggested by the respective authors.

For the BOBYQA based algorithm [103], from Figure 4.3 we observed that the loss function is influenced the most by the pixels in high-contrast areas. Hence, we first apply the variance sub-sampling method followed by block lifting as described in Section 4.2.2[5]. Here, we consider an initial domain of dimension $n_1 = 2 \times 2 \times 3$, and double the refinement of the grid at each layer, i.e. $n_{\ell+1} = 4n_\ell$. Moreover, we observe for (4.6), the choice of a linear model to approximate the loss function works best, and we consequently consider the linear approximation in this chapter; i.e., $\mathbf{M} = \mathbf{0}$ and $q = n + 1$ at all iterations, see [103]. The BOBYQA based algorithm is summarized in Algorithm 5 and a Python implementation of the proposed algorithm based on BOBYQA package from [10] is available on Github[6].

[4]GenAttack: `https://github.com/nesl/adversarial_genattack`
Parsimonious algorithm: `https://github.com/snu-mllab/parsimonious-blackbox-attack`
Square algorithm: `https://github.com/max-andr/square-attack`
Frank-Wolfe algorithm `https://github.com/uclaml/Frank-Wolfe-AdvML`

[5]The choice for this kind of lifting was driven by preliminary experiments in which we considered also a grid method with linear interpolation and a random lifting method as well. It is possible to run the analysis thanks to the code in [4]

[6]`https://github.com/giughi/An-Empirical-Study-of-DFO-Algorithms-for-Targeted-Black-Box-Attacks`

Algorithm 5: BOBYQA Based Algorithm

Input: Image $\mathbf{X} \in \mathbb{R}^n$, target label t, maximum perturbation ε_∞, Neural Net F, initial hierarchical level grid dimensions m, maximum number of queries n^{max}, batch sampling size b, and maximum number κ of queries that we are allowed to do for each batch.
Initialize $\boldsymbol{\eta} \leftarrow \underline{0} \in \mathbb{R}^n$, $n_{eval} = 0$, $\ell = 1$, $n_\ell = 12$.
while $\arg\max F(\mathbf{X} + \boldsymbol{\eta}) \neq t$ and $n_{eval} < n^{max}$ **do**
    # Compute the number of sub samplings necessary to cover the whole domain
    $num_{sub} = n/(n_\ell * b)$
    # Generate the lifting matrix
    $\mathbf{D}_\ell$ = GENERATE_LIFTING(n_ℓ, n)
    # Minimize on all the sampled sub-domains
    for $j = 1, \ldots, num_{sub}$ **do**
        # Compute the matrix which selects b dimensions of the m-dimensional domain.
        $\boldsymbol{\Omega}_\ell^j$ = GENERATE_SAMPLING_MATRIX$(\mathbf{X} + \boldsymbol{\eta}, n_\ell, b, j)$
        # Define the pixel-wise bounds for a perturbation over $\boldsymbol{X} + \boldsymbol{\eta}$.
        $\mathbf{a} = \min\{l - \boldsymbol{\eta}, 0\}, \quad \mathbf{b} = \max\{u - \boldsymbol{\eta}, 0\}$
        # Find $\hat{\boldsymbol{\eta}}_\ell^j$ *by implementing the BOBYQA optimization to the problem (4.9).*
        $\hat{\boldsymbol{\eta}}_\ell^j$=BOBYQA$(F, \mathbf{X}, \boldsymbol{\eta}, \mathbf{a}, \mathbf{b}, \mathbf{D}_\ell, \boldsymbol{\Omega}_\ell^j, t)$ # Algorithm 6
        # Update the noise
        $\boldsymbol{\eta} += \mathbf{D}_\ell \boldsymbol{\Omega}_\ell^j \hat{\boldsymbol{\eta}}_\ell^j$.
        n_{eval} += κ.
    end for
    $\ell += 1$, $n_\ell * = 4$.
end while
if $\arg\max F(\mathbf{X} + \boldsymbol{\eta}) = t$ **then**
    The perturbation is successful.
else if $n_{eval} > n^{max}$ **then**
    The perturbation was not successful with n^{max} iterations.
end if

Algorithm 6: BOBYQA(F, $\mathbf{X}$, $\boldsymbol{\eta}$, $\mathbf{a}$, $\mathbf{b}$, $\boldsymbol{\Omega}_\ell^j$, $\mathbf{D}^\ell$, t, κ)

1: Consider the restricted loss function $\mathcal{L}(\mathbf{X}+\boldsymbol{\eta}, \mathbf{D}^\ell\boldsymbol{\Omega}_\ell^j(\cdot)) : \mathbb{R}^b \to \mathbb{R}$
2: Build an initial model m_0 as in (4.3) of the loss function based on $b+1$ samples; the samples consist of the initial perturbation $\mathbf{X}+\boldsymbol{\eta}$ and the b perturbations obtained by considering changes along the canonical directions of $\mathbf{x}$ in $\mathbf{X}+\boldsymbol{\eta}+\mathbf{D}^\ell\boldsymbol{\Omega}_\ell^j\mathbf{x}$.
3: Find minimizer $\mathbf{x}$ of m_o such that $\mathbf{D}^\ell\boldsymbol{\Omega}_\ell^j\mathbf{x} \in [\mathbf{a}, \mathbf{b}]$.
4: **for** $j = 1, \ldots, \kappa - b$ **do**
5: Add $\mathbf{x}$ to the set of samples and get rid of the least informative one according to [80].
6: Build the new model m_j according to (4.6).
7: Find minimizer $\mathbf{x}$ of m_j such that $\mathbf{D}^\ell\boldsymbol{\Omega}_\ell^j\mathbf{x} \in [\mathbf{a}, \mathbf{b}]$.
8: **end for**
9: Return $\mathbf{x}$.

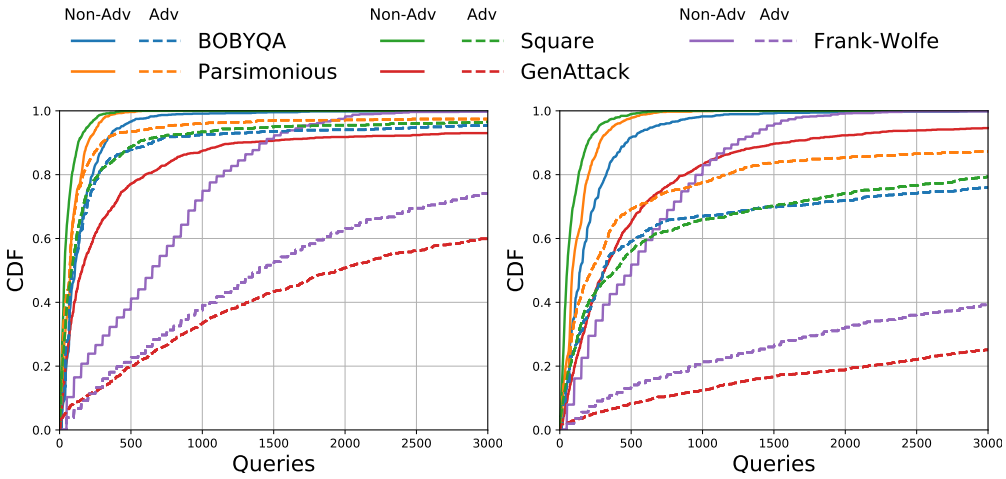
4.3.2 Dataset and Neural Network Specifications

We performed experiments using the popular ResNet50 architecture [38] with two training scenarios; one with the unperturbed images, and one with the defense[7] proposed in [24]. The number of experiments and the choice of the targets for each individual dataset is described below.

CIFAR10 The CIFAR10 dataset contains images from 10 classes and of dimension 32x32x3. To generate a comprehensive distribution for the queries at each energy budget, ten correctly classified images are consider per each class, and each of them is targeted to all of the 9 remaining classes; this way we generate a total of 900 attacks per maximum perturbation energy per adversarial method.

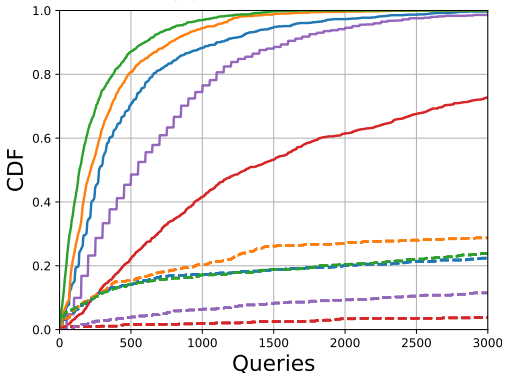
ImageNet This dataset contains millions of images with a dimension of 299x299x3 divided among 1000 classes. Because of the high dimensionality and number of classes, random images are attacked considering a random target class. We conducted 200 and 160 tests for neural networks trained both with and without adversarial training per maximum perturbation energy.

[7]These networks are available already trained at `https://github.com/MadryLab/robustness`

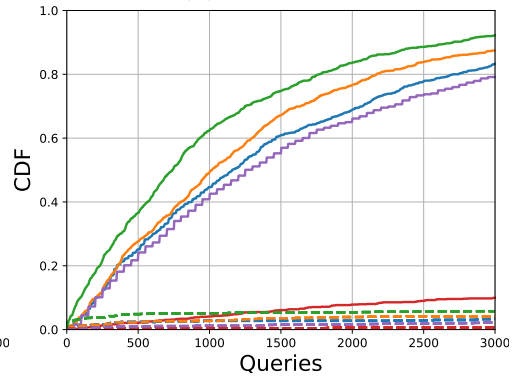


(a) $\varepsilon_\infty = 0.15$

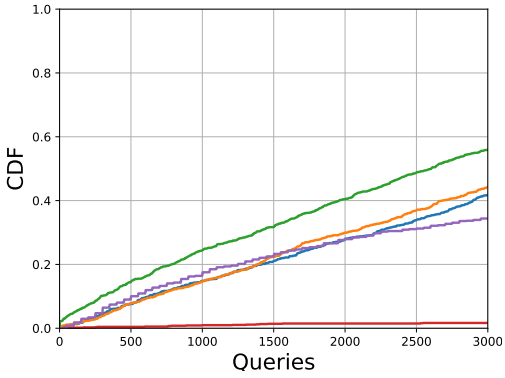
(b) $\varepsilon_\infty = 0.1$



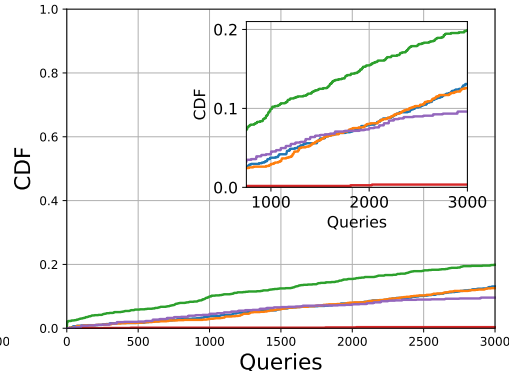
(c) $\varepsilon_\infty = 0.05$



(d) $\varepsilon_\infty = 0.02$

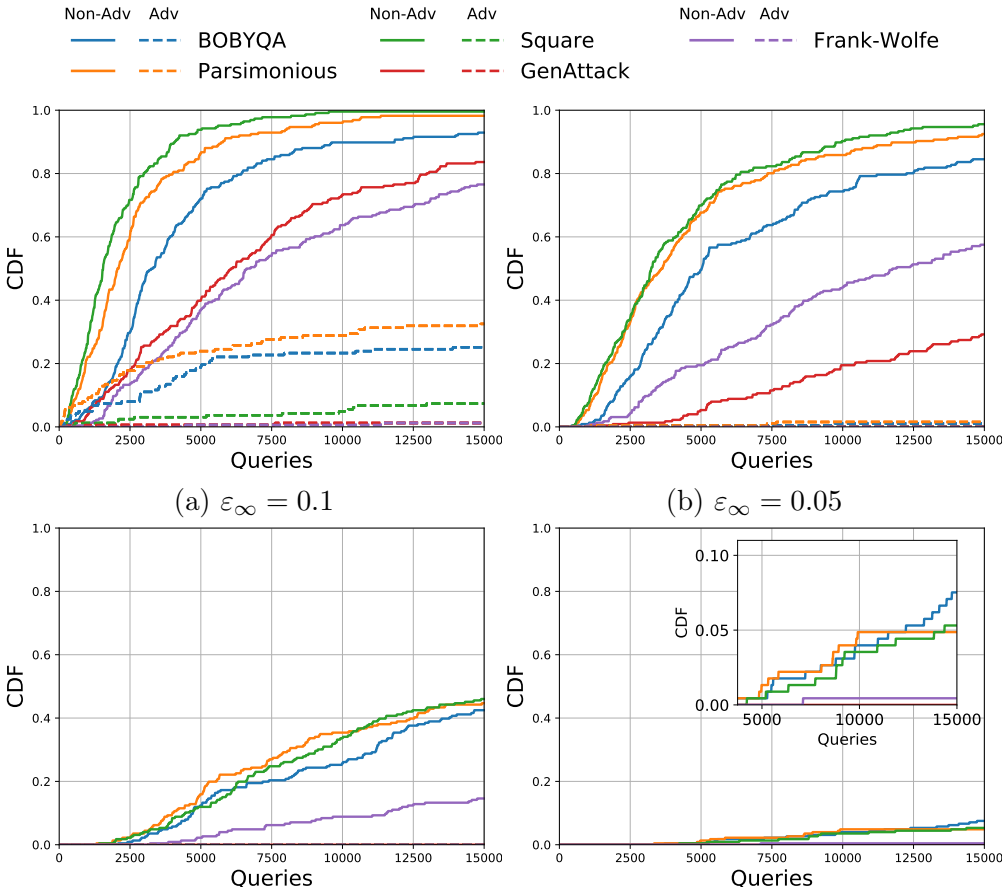


(e) $\varepsilon_\infty = 0.01$



(f) $\varepsilon_\infty = 0.005$

Figure 4.7: Cumulative fraction of test set images successfully misclassified with adversarial examples generated by GenAttack, Parsimonious, Square, Frank-Wolfe, and our BOBYQA based approaches for different maximum perturbation energies ε_∞ and neural networks trained on the CIFAR10 dataset. In all results the solid and dashed lines denoted by 'Non-Adv' and 'Adv' corresponds to attacks on networks trained without or with the MadryLab defense strategy [24] respectively.



(a) $\varepsilon_\infty = 0.1$

(b) $\varepsilon_\infty = 0.05$

(c) $\varepsilon_\infty = 0.02$

(d) $\varepsilon_\infty = 0.01$

Figure 4.8: Cumulative fraction of test set images successfully misclassified with adversarial examples generated by GenAttack, Parsimonious, Square, Frank-Wolfe, and our BOBYQA based approaches for different maximum perturbation energies ε_∞ and neural networks trained on the ImageNet dataset. In all results the solid and dashed lines denoted by 'Non-Adv' and 'Adv' corresponds to attacks on neural networks trained without or with the MadryLab defense strategy [24] respectively.

4.3.3 Results for Standard and MadryLab Trained Neural Networks

In Figures 4.7 and 4.8 we present the CDF for different maximum perturbation energies ε_∞. The pixels are normalized to be in the interval $(-1/2, 1/2)$, hence, $\varepsilon_\infty = 0.1$ would imply that any pixel is allowed to change 10% of the total intensity range from its initial value. The CDFs are illustrated so that we can easily see which method has been able to misclassify the largest fraction of images in the given test-set for a fixed number of queries to the neural network. The confidence intervals of the CDFs are reported in Appendix B and they entail that the CDFs identify the best algorithms in the CIFAR10 case almost surely while in the ImageNet one with high confidence.

For the CIFAR10 dataset in Figure 4.7, we observe that algorithms that search the perturbation directly in the vertices of the perturbation domain require the least amount of neural network queries. In the case of non-adversarially trained neural networks, the Square algorithm is able to misclassify using the least number of queries; this is demonstrated by its associated solid green CDF being consistently above that of the other methods. Specifically, when $\varepsilon_\infty = 0.05$, at 1,000 queries Square algorithms has a CDF of 0.97 compared to 0.94 and 0.88 of the Parsimonious and BOBYQA methods respectively, and for $\varepsilon_\infty = 0.005$ at 3,000 queries Square achieves a CDF of 0.20 which is 50% times higher than Parsimonious and BOBYQA. When the net is instead trained adversarially, dashed lines, Square algorithm looses a lot of its effectiveness becoming comparable to the BOBYQA based method, while Parismonious algorithm achieves almost always the highest fraction of successfully perturbed images for any given maximum number of queries. For example, when $\varepsilon_\infty = 0.05$ at 3,000 queries the CDF of Parisomonious is 0.29 compared to 0.25 and 0.23 of Square and BOBYQA.

In the ImageNet dataset, see Figure 4.8(a), we observe that an adversarial method can be especially susceptible to particular defenses. Specifically, when the neural network is trained without a defense, the Square algorithm has a success rate CDF that is consistently higher than the other methods, but the success rate CDF for the Square algorithm is decreased by the MadryLab defense so that it is substantially less effective than Parsimonious and BOBYQA algorithms. On the

other hand, the Parsimonious method achieves similar results to Square algorithm in the non-adversarial case. On average for the different maximum perturbation energies Parsimonious is 0.045 less efficient than Square, but when the defense is introduced it finds the adversarial examples with the least number of queries. In Figure 4.8(a) Parisomious has a CDF of 0.33 at 15,000 queries while BOBYQA 0.24 and Square 0.07. The rate with which the CDFs decrease as the maximum perturbation energy ε_∞ decreases it also differs by algorithm. The CDF for Square decreases moderately faster than for Parsimonious such that Square has a consistently higher CDF than Parsimonious for $\varepsilon = 0.1$ in Figure 4.8(a) but consistently lower in Figure 4.8(d). Moreover, the success rate for BOBYQA decreases the slowest with ε_∞ such that in Figure 4.8 its CDF is similar to or greater than Parsimonious. Specifically, in Figure 4.8(d) at 15,000 the final CDF of BOBYQA algorithm queries is 1.42 times higher than the one of the Square algorithm.

The Frank-Wolfe algorithm is able to achieve results comparable to the ones of the methods above while considering the small-dimensional problem of CIFAR10 with a very low maximum perturbation energy. However, when considering the ImageNet case and the adversarially trained neural networks, the Frank-Wolfe algorithm has a substantially lower success rate CDF; e.g. in the ImageNet case with non-adversarial training, Square algorithm achieves a CDF 1.66 times higher than the Frank-Wofle algorithm when $\varepsilon_\infty = 0.05$.

GenAttack has a higher success rate CDF than the Frank-Wolfe algorithm in the ImageNet case for $\varepsilon_\infty = 0.1$, see Figure 4.8(a), but, besides this case, it constantly achieves the lowest success rate.

The relative success of the misclassification algorithms as a function of the allowed perturbation energy is determined by the training loss function and associated partition of the input space into regions associated with each class [53][Figure 3]. Each correctly classified example, not on the boundary between classes, has a small enough ε_∞ region surrounding it where misclassification cannot be obtained. As ε_∞ is increased the fraction of the perturbations which admit misclassification can be expected to increase, and misclassification becomes trivial once ε_∞ is sufficiently large that the majority of vertices correspond to misclassification. In fact, [32] suggest the pre-classification output vector is maximally misclassified according to (4.2) at vertices.

For these reasons we can expect that vertex search methods such as Parsimonious and Square are preferable for large ε_∞ while model based methods such as BOBYQA and Frank-Wolfe become increasingly beneficial, relatively, as ε_∞ decreases and the fraction of vertices which correspond to misclassification becomes small. Although both BOBYQA and Frank-Wolfe are based on a linear approximation of the problem, their respective number of samples taken differs substantially by how the samples are selected to construct the model and the subspaces with which they optimize over. In particular, Frank-Wolfe optimizes over $b = 25$ dimensional subspaces drawn at random from the unit sphere while BOBYQA sequentially optimizes over subsampled batches of variables as described in Section 4.2.1. The relative impact of these differing dimensionality reduction techniques has not been explored and may account for some of the superior performance of BOBYQA as compared to Frank-Wolfe.

4.3.4 Results with Fixed Pixel Count Constraints

In addition to neural network training designed to increase robustness, such as MadryLab considered previously, there are a multitude of other defenses and real world constraints [37]. The relative success rate, or other characteristics, of adversarial algorithms can be expected to differ in these diverse settings. To demonstrate this, we consider one such setting where the maximum number of pixels allowed to be perturbed is limited. This is motivated by the defenses where neural network inputs are thresholded in a wavelet domain to exclude high frequency perturbations [36], as well as by real world constraints such as attacks designed to appear structured such as localized perturbations designed to look like graffiti [25, 64]. We allow the algorithms to perturb only the fixed selection of the 1,000 pixels of the targeted image that have the highest variance in intensity in their channel neighborhood. Because of the previous results it is possible to identify three methods that work consistently better than the others, and thus only these will be considered, namely: the Parsimonious, the Square, and the BOBYQA based algorithms. To allow the perturbations to be limited to the selected pixels, we consider the Square algorithm with squares of pixel dimension, the Parsimonious

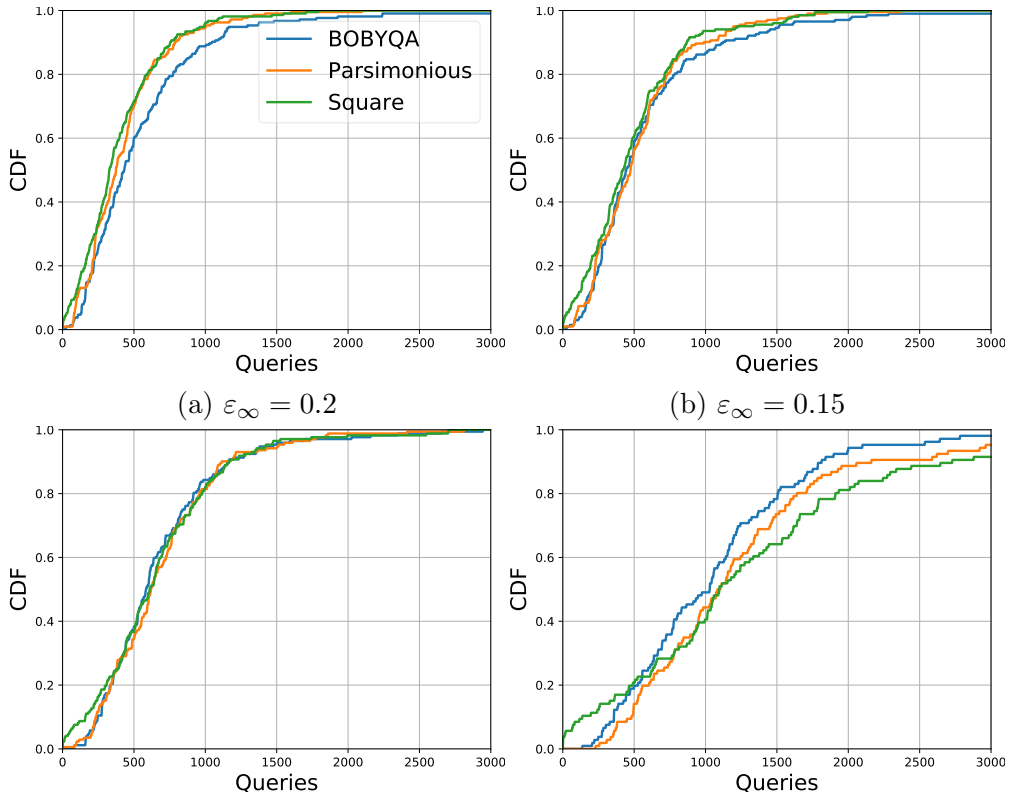
algorithm on the finest grid, and the BOBYQA algorithm without the hierarchical lifting, i.e. $\mathbf{D}^1 = \mathbf{I}$ where $\mathbf{I}$ is the identity matrix.

The results reported in Figure 4.9 suggest that when the domain is dimensionally limited, the most efficient algorithm changes according to the allowed maximum perturbation energy. When the maximum perturbation energy decreases and the linear model is more accurate, the BOBYQA method manages to achieve a higher SR than both Square and Parsimonious algorithms, unlike in the previous experiments. Moreover, the Parsimonious algorithm has almost identical behavior to Square algorithm for high energy bounds, but becomes more efficient when the maximum energy is $\varepsilon_\infty = 0.05$. Figures 4.9 and 4.7 also differ by the former not employing hierarchical lifting as described in Section 4.2.2 while Figure 4.7 does make use of lifting. The overall trends between Figures 4.7 and 4.9 are consistent which suggests that the use of lifting does not change the overall trends observed between classes of methods, rather it consistently reduces the overall number of samples needed for misclassification.

We also considered experiments on ImageNet, but limiting the number of pixels that could be perturbed did not allow for any successful misclassification with less than 15,000 queries.

4.3.5 Relative Computational Cost

While the focus in this manuscript is to compare the different typology of algorithms according to the their misclassification success rate as a function of the number of queries to the neural network; it is also worth noting that the different type of DFO algorithms can be expected to have differing computational burdens. Table 4.1 displays the average time for each algorithm to update their perturbation of the input per 1,000 queries. In particular, these results are obtained by running 10 attacks to the ResNet50 non-adversarially trained on ImageNet with a perturbation error of $\varepsilon_\infty = 10^{-3}$, and the time was then averaged on one thousand queries. However, we remark that these algorithms were not optimised on a computational point of view and these results are reported mainly for an indicative purpose.



(a) $\varepsilon_\infty = 0.2$ (b) $\varepsilon_\infty = 0.15$

(c) $\varepsilon_\infty = 0.1$ (d) $\varepsilon_\infty = 0.05$

Figure 4.9: Cumulative fraction of test set images successfully misclassified with adversarial examples generated by Parsimonious, Square, and our BOBYQA based approaches for different maximum perturbation energies ε_∞ against a ResNet50 trained non-adversarially on the CIFAR10 dataset when only the 1000 pixels with the highest variance in intensity in their neighborhood are allowed to be modified.

All the algorithms have computational costs on the same order of magnitude. The Square algorithm stands out as having the lowest computational burden and achieving state-of-the-art misclassification rates for non-adversarially trained neural networks. However, for neural networks trained with the MadryLab defense, Parsimonious is observed to have a superior misclassification rate, at the cost of taking approximately 4 times longer to compute the perturbation. Finally, the fact that BOBYQA and Parsimonious algorithms are the slowest shows how sophisticated hierarchical approaches are computationally intensive, though can be beneficial for lower perturbation energies or neural networks trained against adversarial attacks.

	BOBYQA	GenAttack	Frank-Wolfe	Parismonious	Square
Time/1'000 queries	43.7s	11.7s	18.6s	51.8s	12.7s

Table 4.1: Average time required by different algorithms in processing 1'000 queries to the ImageNet non-adversarially trained ResNet50.

4.4 Discussion and Conclusion

We have compared for the first time how the the existing GenAttack [4], Parsimonious [60], Square [5], and Frank-Wolfe [12] algorithms, and the newly introduced BOBYQA based method, behave when the available ℓ^∞ energy for a perturbation varies, and an adversarial training or a structural defense is considered.

The results suggest that those methods limiting the search for an adversarial example to the vertices of the ℓ^∞ perturbation domain generally work better. Whilst Square algorithm is especially effective on the non-adversarially trained neural networks, the Parsimonious algorithm manages to outperform any other approach when the neural networks are adversarially trained with the MadryLab implementation. Furthermore, the Parsimonious algorithm performs better than Square when considering the structural defense that limits the attacks on some pixels, suggesting that an algorithm based on combinatorial search is robust in its hyper-parameters to the setting where it is applied.

The BOBYQA based algorithm was introduced in this chapter to explore how model-based approaches compare to the state-of-the-art algorithms, and was found to achieve similar results to the Parsimonious and Square algorithms. In almost in all the experiments the BOBYQA based algorithm achieves a success rate CDF comparable to the ones of the Parsimonious and the Square algorithms; it achieves the state-of-the-art success rate at saturation for low maximum perturbation energy constraint both in the ImageNet case and in the pixel constrained problem, thus becoming the preferable choice in these cases. Figures 4.7 and 4.9 differ in part by the former making use of hierarchical lifting as described in Section 4.2.2 while the later does not employ any lifting. The aforementioned performance trends are consistent in Figures 4.7 and 4.9 which suggests that while lifting reduces the overall number of samples, it does not impact the relative performance between algorithms. New dimensionality reduction techniques are a topic of recent

investigation for DFO, see for example [11], or variations of the derivative-based multi-level approach [34], and might further improve the results observed here.

In conclusion, we find that both the structure of the algorithm and the attack setting have the potential to impact the algorithm performance. These observations highlight the importance of comparing any new algorithm to the state-of-the-art in a variety of different settings, such as is done here. Similarly, the effectiveness of an adversarial defense for neural networks should always be tested using as wide a range of algorithms as possible.

Chapter 5

Conclusion and Future Work

This thesis illustrates the work that I have conducted under the supervision of Prof. Jared Tanner whilst being a PhD student in the 4th cohort of the Industrially Focused Mathematical Modelling program.

The overarching aim of this thesis is the exploration of topics in the field of deep learning with mathematical techniques. Specifically, we focused on two different research directions: the study of the information flow in a random neural network relying on information theory and random matrix theory, and the generation of adversarial examples in the black-box adversarial setting with derivative-free-optimisation techniques.

5.1 Information Propagation in a Random Neural Network

In Chapters 2 and 3 we have analysed how information propagates through random neural networks. In [1] it was observed that according to some estimations of the mutual information at initialisation, there are some non-trivial choices of the initialisation parameters for which the information propagates easily through the layers. We complemented this observation theoretically.

In Chapter 2 we introduced a lower bound for the mutual information and we proposed its approximation by relying on the mean-field approximation. The empirical results that we observed confirmed the observations in [1] of the mutual information being optimal for certain specific parameters. The approximation

that we introduced qualitatively reproduced this behaviour but had a quantitative discrepancy with the empirical measure.

To improve the approximation, in Chapter 3 we modelled the same lower bound by considering the eigen-distribution of the matrices that determine the bound. To do this, we expanded the work done in [72] by improving the accuracy of the distribution estimation for deeper layers, and by also considering a different matrix. These analyses resulted in an improvement of the approximation at early layers, once again confirming the observation in [1]. However, the computational complexity of this new approximation does not allow an agile computation of the lower bound for deep layers. Nevertheless, its formulation allows us to identify the variables that affect the information flow decay.

5.2 Adversarial Examples

In Chapter 4 we focused on exploring how different derivative-free optimisation methods perform when applied to the generation of adversarial examples in the black-box setting.

We have compared for the first time how the existing adversarial algorithms behave when the available ℓ^∞ energy for a perturbation varies, and an adversarial training or a structural defence is considered. The results suggest that those methods limiting the search for an adversarial example to the vertices of the ℓ^∞ perturbation domain generally work better and that an algorithm based on combinatorial search is robust in its hyper-parameters to the setting where it is applied.

The BOBYQA based algorithm was introduced to explore how model-based approaches compare to the state-of-the-art algorithms, and was found to achieve similar results.

In conclusion, we find that both the structure of the algorithm and the attack setting have the potential to impact the algorithm performance. These observations highlight the importance of comparing any new algorithm to the state-of-the-art in a variety of different settings, such as is done here. Similarly, the effectiveness of an adversarial defence for DNNs should always be tested using as wide a range of algorithms as possible.

5.3 Future Work

There are many directions in which we can imagine expanding the work that was presented in this thesis.

With respect to the study of mutual information propagation, we could consider expanding the work done relying on the Stieltjes transform in Chapter 3 to the case where $\sigma_b \neq 0$. The work in [2] did model the distribution of $\mathbf{\Lambda}_{\boldsymbol{h}^{(\ell)}}$ when $\sigma_b > 0$, but it relied on the resolvent method rather than on the method of moments. If the same analysis was applicable to $\mathbf{\Lambda}_{h^l} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}_{xh^l}^\top \mathbf{\Sigma}_{xh^l}$, it would be possible to understand if any further variable affects the decay of the mutual information at initialisation. This would allow us to have a complete picture of the relation between the choice of $(\sigma_w, \sigma_b, \phi)$ and the information flow in the neural network, and it could lead to improvements in the design of activation functions. Furthermore, it would be interesting to explore empirically if an initialisation, where the weights are samples from an optimal distribution, do actually lead to any improvements in the training time and quality of the solution.

It would also be interesting to understand if it is possible to determine how tight the lower bound we have defined is. For example, the results in [55] would allow to define the upper bound considering the joint distribution of the inputs and hidden layers being log-concave. However, the convergence of neural networks to Gaussian processes when the width goes to infinity might make this bound redundant.

Finally, relative to the adversarial examples generation, more theoretical work should be considered to improve the efficacy of the DFO based adversarial methods. New dimensionality reduction techniques are a topic of recent investigation for DFO, see for example [11], or variations of the derivative-based multi-level approach [34], and thanks to their convergence analysis, the choice of the hyper-parameters in the adversarial methods could be driven by expected improvements rather than heuristics. It would also be interesting to study if the results that we observed in our controlled setting are descriptive of how the different methods perform in practice when applied to real world problems therefore validating the need of more theoretically principled methods.

Appendix A

Proofs Relevant to the Method of Moments Applied to $\mathbf{\Lambda}_{h^{(\ell)}}$ and $\mathbf{\Lambda}_{h^{(\ell)}} - \frac{1}{\sigma_x^2}\mathbf{\Sigma}^{\top}_{xh^{(\ell)}}\mathbf{\Sigma}_{xh^{(\ell)}}$

A.1 Proofs Relevant to Theorem 3.2.2

A.1.1 Initial Result

The expectation of any $2k$ cycle can be expressed in a different format thanks to Lemma A.1.1. From such a formulation it is then possible to identify the expected contribution for $k = 1$ and for $k > 1$.

Lemma A.1.1. *For* $\boldsymbol{M} = \frac{1}{m}\boldsymbol{Y}\boldsymbol{Y}^{\top}$ *and* $\boldsymbol{Y} = \phi\left(\boldsymbol{W}\tilde{\boldsymbol{X}}\right)$ *as defined in Theorem 3.2.2, then*

$$E_{2k}^{(n)} = \mathbb{E}\left[\boldsymbol{Y}_{i_1\mu_2}\boldsymbol{Y}_{i_2\mu_1}\ldots\boldsymbol{Y}_{\mu_k i_1}\right] \tag{A.1}$$

$$= \int \left(\sum_{\nu=0}^{\infty} \frac{\left(-\frac{1}{2}\sum_{\xi\geq 2}\frac{(-1)^{\xi+1}}{\xi} tr\mathbf{\Sigma}^{\xi} tr\left(\left(\bar{\mathbf{\Lambda}}\bar{\mathbf{\Lambda}}^{\top}\right)^{\xi}\right)\right)^{\nu}}{\nu!}\right) \cdot \left(\prod_{\lambda_{\alpha\beta}\in\mathcal{Z}} e^{-\frac{tr\Sigma}{2}\bar{\lambda}_{\alpha\beta}^2 - i\,\frac{\sqrt{n}}{\sigma_w\sigma_x}\bar{\lambda}_{\alpha\beta}z_{\alpha\beta}}\right) F(z)\mathcal{D}z\mathcal{D}\mathbf{\Sigma}\mathcal{D}\bar{\lambda} \tag{A.2}$$

where $\mathcal{Z}$ *is the set of combinations* $\{(i_1,\mu_1),(i_2,\mu_1),\ldots,(i_1,\mu_k)\}$ *on which the*

function $F(z) = \prod_{(\alpha\beta)\in\mathcal{Z}} \phi(z_{\alpha\beta})$*, the variables*

$$\boldsymbol{Z}_{i\mu} = \begin{cases} z_{i\mu} & \text{if } (i,\mu) \in \mathcal{Z} \\ 0 & \text{otherwise.} \end{cases} \qquad \text{and} \qquad \bar{\boldsymbol{\Lambda}}_{i\mu} = \begin{cases} \frac{\sigma_x \sigma_w}{\sqrt{n}} \bar{\lambda}_{i\mu} & \text{if } (i,\mu) \in \mathcal{Z} \\ 0 & \text{otherwise.} \end{cases} \tag{A.3}$$

and the measures

$$\mathcal{D}z = \prod_{z_{\alpha\beta}\in\mathcal{Z}} dz_{\alpha\beta} \qquad \text{and} \qquad \mathcal{D}\bar{\lambda} = \prod_{\lambda_{\alpha\beta}\in\mathcal{Z}} \frac{d\lambda_{\alpha\beta}}{2\pi\sigma_x\sigma_w/\sqrt{n}} \tag{A.4}$$

are defined.

Proof. Now we consider auxiliary integrals over z, by adding delta functions enforcing $\mathbf{Z} = \mathbf{W}\boldsymbol{\Sigma}^{1/2}\ \mathbf{X}$ with

$$\mathbf{Z}_{i\mu} = \begin{cases} z_{i\mu} & \text{if } (i,\mu) \in \mathcal{Z} \\ 0 & \text{otherwise.} \end{cases} \tag{A.5}$$

where $\mathcal{Z}$ denotes the set of unique pairs (i,μ) in equation (3.59):

$$E_{2k}^{(n)} = \int \prod_{(\alpha,\beta)\in\mathcal{Z}} \delta(z_{\alpha\beta} - \sum_k \mathbf{W}_{\alpha k}\boldsymbol{\Sigma}_{kk}^{1/2}\mathbf{X}_{k\beta})\phi(z_{i_1\mu_1})\phi(z_{i_2\mu_1})...\phi(z_{i_1\mu_k})\mathcal{D}z\mathcal{D}\mathbf{W}\mathcal{D}\boldsymbol{\Sigma}\mathcal{D}\mathbf{X} \tag{A.6}$$

where

$$\mathcal{D}z = \prod_{(\alpha,\beta)\in\mathcal{Z}} dz_{\alpha\beta}. \tag{A.7}$$

Now we consider the Fourier expression of the Dirac δ

$$\delta(x) = \frac{1}{2\pi}\int e^{i\lambda x} d\lambda \tag{A.8}$$

and therefore introduced the matrix $\boldsymbol{\Lambda} \in \mathbb{R}^{n\times m}$ whose entries are

$$\boldsymbol{\Lambda}_{i\mu} = \begin{cases} \lambda_{i\mu} & \text{if } (i,\mu) \in \mathcal{Z} \\ 0 & \text{otherwise.} \end{cases} \tag{A.9}$$

with

$$\mathcal{D}\lambda = \prod_{(\alpha,\beta)\in\mathcal{Z}} \frac{d\lambda_{\alpha\beta}}{2\pi}. \tag{A.10}$$

to obtain

$$E_{2k}^{(n)} = \int \prod_{z_{\alpha\beta}\in\mathcal{Z}} \delta(z_{\alpha\beta} - \sum_k \mathbf{W}_{\alpha k}\mathbf{\Sigma}_{kk}^{1/2}\mathbf{X}_{k\beta})\phi(z_{i_1\mu_1})\phi(z_{i_2\mu_1})...\phi(z_{i_1\mu_k})\mathcal{D}z\mathcal{D}\mathbf{W}\mathcal{D}\mathbf{\Sigma}\mathcal{D}\mathbf{X}$$

(A.11)

$$= \int \prod_{z_{\alpha\beta}\in\mathcal{Z}} exp\left(-i\lambda_{\alpha\beta}\left(\sum_k \mathbf{W}_{\alpha k}\mathbf{\Sigma}_{kk}^{1/2}\mathbf{X}_{k\beta} - z_{\alpha\beta}\right)\right)\cdot$$
$$\cdot\, \phi(z_{i_1\mu_1})\phi(z_{i_2\mu_1})...\phi(z_{i_1\mu_k})\mathcal{D}z\mathcal{D}\mathbf{W}\mathcal{D}\mathbf{\Sigma}\mathcal{D}\mathbf{X}\mathcal{D}\lambda \tag{A.12}$$

$$= \int exp\left(-i\sum_{z_{\alpha\beta}\in\mathcal{Z}} \lambda_{\alpha\beta}\underbrace{\left(\sum_k \mathbf{W}_{\alpha k}\mathbf{\Sigma}_{kk}^{1/2}\mathbf{X}_{k\beta} - z_{\alpha\beta}\right)}_{(\mathbf{W}\mathbf{\Sigma}^{1/2}\mathbf{X}-\mathbf{Z})_{\alpha\beta}}\right)\cdot$$
$$\cdot\, \phi(z_{i_1\mu_1})\phi(z_{i_2\mu_1})...\phi(z_{i_1\mu_k})\mathcal{D}z\mathcal{D}\mathbf{W}\mathcal{D}\mathbf{\Sigma}\mathcal{D}\mathbf{X}\mathcal{D}\lambda \tag{A.13}$$

$$= \int exp\left(-i\sum_{\alpha,\beta=1}^{n,m} \mathbf{\Lambda}_{\alpha\beta}(\mathbf{W}\mathbf{\Sigma}^{1/2}\mathbf{X} - \mathbf{Z})_{\alpha\beta}\right)\cdot$$
$$\cdot\, \phi(z_{i_1\mu_1})\phi(z_{i_2\mu_1})...\phi(z_{i_1\mu_k})\mathcal{D}z\mathcal{D}\mathbf{W}\mathcal{D}\mathbf{\Sigma}\mathcal{D}\mathbf{X}\mathcal{D}\lambda \tag{A.14}$$

$$= \int e^{-i\ tr\left(\Lambda^\top(\mathbf{W}\mathbf{\Sigma}^{1/2}\mathbf{X}-\mathbf{Z})\right)}\phi(z_{i_1\mu_1})\phi(z_{i_2\mu_1})...\phi(z_{i_1\mu_k})\mathcal{D}z\mathcal{D}\mathbf{W}\mathcal{D}\mathbf{\Sigma}\mathcal{D}\mathbf{X}\mathcal{D}\lambda$$

(A.15)

where $tr()$ corresponds to the trace function.

Now we first integrate over $\mathbf{X}$ the factors of (A.15) that depend on it

$$\int e^{-itr(\mathbf{\Lambda}^\top\mathbf{W}\mathbf{\Sigma}^{1/2}\mathbf{X})}\mathcal{D}\mathbf{X} = \prod_{b,c=1}^{m,n}\int \frac{d\mathbf{X}_{cb}}{\sqrt{2\pi\sigma_x^2}} exp\left[-\frac{1}{2\sigma_x^2}\mathbf{X}_{cb}^2 - i\sum_{a=1}^{n}\lambda_{ab}\mathbf{W}_{ac}\mathbf{\Sigma}_{cc}^{1/2}\mathbf{X}_{cb}\right]$$
$$= \prod_{b,c=1}^{m,n}\int \frac{d\mathbf{X}_{cb}}{\sqrt{2\pi\sigma_x^2}} exp\left[-\frac{1}{2\sigma_x^2}\left(\mathbf{X}_{cb} + i\sigma_x^2(\sum_{a=1}^{n}\lambda_{ab}\mathbf{W}_{ac}\mathbf{\Sigma}_{cc}^{1/2})\right)^2 - \frac{\sigma_x^2}{2}(\sum_{a=1}^{n}\lambda_{ab}\mathbf{W}_{ac}\mathbf{\Sigma}_{cc}^{1/2})^2\right] \tag{A.16}$$
$$= \prod_{b,c=1}^{m,n} exp\left[-\frac{\sigma_x^2}{2}(\sum_{a=1}^{n}\lambda_{ab}\mathbf{W}_{ac}\mathbf{\Sigma}_{cc}^{1/2})^2\right]$$

$$\underbrace{\int \frac{d\mathbf{X}_{cb}}{\sqrt{2\pi\sigma_x^2}} exp\left[-\frac{1}{2\sigma_x^2}\left(\mathbf{X}_{cb}+i\sigma_x^2(\sum_{a=1}^{n}\lambda_{ab}\mathbf{W}_{ac}\mathbf{\Sigma}_{cc}^{1/2})\right)^2\right]}_{=1} \tag{A.17}$$

$$= \prod_{b,c=1}^{m,n} exp\left[-\frac{\sigma_x^2}{2}(\sum_{a=1}^{n}\lambda_{ab}\mathbf{W}_{ac}\mathbf{\Sigma}_{cc}^{1/2})^2\right] = exp\left[-\frac{\sigma_x^2}{2}\sum_{b,c=1}^{m,n}(\sum_{a=1}^{n}\lambda_{ab}\mathbf{W}_{ac}\mathbf{\Sigma}_{cc}^{1/2})^2\right] \tag{A.18}$$

$$= exp\left[-\frac{\sigma_x^2}{2}\|\mathbf{\Lambda}^\top\mathbf{W}\mathbf{\Sigma}^{1/2}\|_F^2\right] = e^{-\frac{\sigma_x^2}{2}tr(\mathbf{\Sigma}^{1/2\top}\mathbf{W}^\top\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{W}\mathbf{\Sigma}^{1/2})} \tag{A.19}$$

where in equation (A.17) we used the property the complex integral of $z = x + iy$ over the closed cycle $(-\infty, \infty, i\mu + \infty, i\mu - \infty)$ of the analytical function $\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-iy)^2/(2\sigma^2)}$ is null and therefore

$$\int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x-i\mu)^2/(2\sigma^2)}dx = \int_{-\infty}^{\infty}\frac{1}{\sqrt{2\pi\sigma^2}}e^{-(x)^2/(2\sigma^2)}dx = 1.$$

Now we integrate over $\mathbf{W}$

$$\int e^{-\frac{\sigma_x^2}{2}tr(\mathbf{\Sigma}^{1/2\top}\mathbf{W}^\top\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{W}\mathbf{\Sigma}^{1/2})}\mathcal{D}\mathbf{W}$$

$$= \int\left(\prod_{i,j=1}^{n}\frac{d\mathbf{W}_{ij}}{\sqrt{2\pi\sigma_w^2/n}}e^{-\frac{n\mathbf{W}_{ij}^2}{2\sigma_w^2}}\right)e^{-\frac{\sigma_x^2}{2}tr(\mathbf{\Sigma}^{1/2\top}\mathbf{W}^\top\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{W}\mathbf{\Sigma}^{1/2})} \tag{A.20}$$

$$= \int\left(\prod_{i,j=1}^{n}\frac{d\mathbf{W}_{ij}}{\sqrt{2\pi\sigma_w^2/n}}\right)e^{-\frac{n\sum_{i,j=1}^{n}\mathbf{W}_{ji}\mathbf{W}_{ij}}{2\sigma_w^2}}e^{-\frac{\sigma_x^2}{2}tr(\mathbf{\Sigma}^{1/2\top}\mathbf{W}^\top\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{W}\mathbf{\Sigma}^{1/2})} \tag{A.21}$$

$$= \int\left(\prod_{i,j=1}^{n}\frac{d\mathbf{W}_{ij}}{\sqrt{2\pi\sigma_w^2/n}}\right)e^{-\frac{ntr\mathbf{W}^\top\mathbf{W}}{2\sigma_w^2}}e^{-\frac{\sigma_x^2}{2}tr(\mathbf{\Sigma}^{1/2\top}\mathbf{W}^\top\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{W}\mathbf{\Sigma}^{1/2})} \tag{A.22}$$

$$= \int\left(\prod_{i,j=1}^{n}\frac{d\mathbf{W}_{ij}}{\sqrt{2\pi\sigma_w^2/n}}\right)e^{-\frac{ntr\mathbf{W}^\top\mathbf{W}}{2\sigma_w^2}-\frac{\sigma_x^2}{2}tr(\mathbf{\Sigma}^{1/2}\mathbf{\Sigma}^{1/2\top}\mathbf{W}^\top\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{W})} \tag{A.23}$$

$$= \int\left(\prod_{j=1}^{n}\frac{d^n\mathbf{W}_j}{(2\pi\sigma_w^2/n)^{n/2}}\right)e^{-\sum_{j=1}^{n}\left(\frac{n}{2\sigma_w^2}\mathbf{w}_j^\top\mathbf{w}_j+\frac{\sigma_x^2}{2}\mathbf{\Sigma}_{jj}\mathbf{w}_j^\top\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{w}_j\right)} \tag{A.24}$$

$$= \int\left(\prod_{j=1}^{n}\frac{d^n\mathbf{W}_j}{(2\pi\sigma_w^2/n)^{n/2}}e^{-\left(\frac{n}{2\sigma_w^2}\mathbf{w}_j^\top\mathbf{w}_j+\frac{\sigma_x^2}{2}\mathbf{\Sigma}_{jj}\mathbf{w}_j^\top\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{w}_j\right)}\right) \tag{A.25}$$

$$= \prod_{j=1}^{n} \int \frac{d^n \mathbf{w}_j}{(2\pi\sigma_w^2/n)^{n/2}} e^{-\frac{1}{2}\mathbf{w}_j^\top \left(\frac{n}{\sigma_w^2}\mathbf{I}+\sigma_x^2\boldsymbol{\Sigma}_{jj}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\right)\mathbf{w}_j} \tag{A.26}$$

$$= \prod_{j=1}^{n} \int \frac{d^n \mathbf{w}_j}{(2\pi\sigma_w^2/n)^{n/2}} \frac{det^{-1}(\mathbf{I}+\frac{\sigma_x^2\sigma_w^2}{n}\boldsymbol{\Sigma}_{jj}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)^{1/2}}{det^{-1}(\mathbf{I}+\frac{\sigma_x^2\sigma_w^2}{n}\boldsymbol{\Sigma}_{jj}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)^{1/2}} e^{-\frac{1}{2}\mathbf{w}_j^\top \left(\frac{n}{\sigma_w^2}\mathbf{I}+\sigma_x^2\boldsymbol{\Sigma}_{jj}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\right)\mathbf{w}_j} \tag{A.27}$$

$$= \prod_{j=1}^{n} \frac{1}{det(\mathbf{I}+\frac{\sigma_x^2\sigma_w^2}{n}\boldsymbol{\Sigma}_{jj}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)^{1/2}} \int \frac{d^n \mathbf{w}_j}{(2\pi\sigma_w^2/n)^{n/2}} \frac{e^{-\frac{1}{2}\mathbf{w}_j^\top \left(\frac{n}{\sigma_w^2}\mathbf{I}+\sigma_x^2\boldsymbol{\Sigma}_{jj}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\right)\mathbf{w}_j}}{det^{-1}(\mathbf{I}+\frac{\sigma_x^2\sigma_w^2}{n}\boldsymbol{\Sigma}_{jj}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)^{1/2}} \tag{A.28}$$

$$= \prod_{j=1}^{n} \frac{1}{det(\mathbf{I}+\frac{\sigma_x^2\sigma_w^2}{n}\boldsymbol{\Sigma}_{jj}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)^{1/2}} \int \frac{d^n \mathbf{w}_j}{(2\pi\sigma_w^2/n)^{n/2}} \frac{e^{-\frac{1}{2}\mathbf{w}_j^\top \left(\frac{n}{\sigma_w^2}\mathbf{I}+\sigma_x^2\boldsymbol{\Sigma}_{jj}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\right)\mathbf{w}_j}}{det\left((\mathbf{I}+\frac{\sigma_x^2\sigma_w^2}{n}\boldsymbol{\Sigma}_{jj}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)^{-1}\right)^{1/2}} \tag{A.29}$$

$$= \prod_{j=1}^{n} \frac{1}{det(\mathbf{I}+\frac{\sigma_x^2\sigma_w^2}{n}\boldsymbol{\Sigma}_{jj}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)^{1/2}} \int \frac{d^n \mathbf{w}_j}{(2\pi)^{n/2}} \frac{e^{-\frac{1}{2}\mathbf{w}_j^\top \left(\frac{n}{\sigma_w^2}\mathbf{I}+\sigma_x^2\boldsymbol{\Sigma}_{jj}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\right)\mathbf{w}_j}}{det\left(\sigma_w^2/n(\mathbf{I}+\frac{\sigma_x^2\sigma_w^2}{n}\boldsymbol{\Sigma}_{jj}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)^{-1}\right)^{1/2}} \tag{A.30}$$

$$= \prod_{j=1}^{n} \frac{1}{det(\mathbf{I}+\frac{\sigma_x^2\sigma_w^2}{n}\boldsymbol{\Sigma}_{jj}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)^{1/2}} \underbrace{\int \frac{d^n \mathbf{w}_j}{(2\pi)^{n/2}} \frac{e^{-\frac{1}{2}\mathbf{w}_j^\top \left(\frac{n}{\sigma_w^2}\mathbf{I}+\sigma_x^2\boldsymbol{\Sigma}_{jj}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top\right)\mathbf{w}_j}}{det\left((\frac{n}{\sigma_w^2}\mathbf{I}+\sigma_x^2\boldsymbol{\Sigma}_{jj}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)^{-1}\right)^{1/2}}}_{=1} \tag{A.31}$$

$$= \prod_{j=1}^{n} \frac{1}{det(\mathbf{I}+\frac{\sigma_x^2\sigma_w^2}{n}\boldsymbol{\Sigma}_{jj}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)^{1/2}} \tag{A.32}$$

where we consider $\mathbf{w}_j$ as the jth column of $\mathbf{W}$ and we used the property that for a general non-singular matrix $det(\mathbf{A}^{-1}) = det^{-1}(\mathbf{A})$.

This implies that by considering $F(z) = \prod_{(\alpha,\beta)\in\mathcal{Z}} \phi(z_{\alpha\beta})$

$$E_{2k}^{(n)} = \int \left(\prod_j \frac{1}{det(\mathbf{I}+\frac{\sigma_x^2\sigma_w^2}{n}\boldsymbol{\Sigma}_{jj}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top)^{1/2}} \right) e^{-i\ tr(\boldsymbol{\Lambda}^\top\mathbf{Z})} F(z)\mathcal{D}z\mathcal{D}\boldsymbol{\Sigma}\mathcal{D}\lambda \tag{A.33}$$

$$= \int exp\left(-\sum_j \frac{1}{2}\log det(\mathbf{I}+\frac{\sigma_x^2\sigma_w^2}{n}\boldsymbol{\Sigma}_{jj}\boldsymbol{\Lambda}\boldsymbol{\Lambda}^\top) - itr\boldsymbol{\Lambda}^\top\mathbf{Z} \right) F(z)\mathcal{D}z\mathcal{D}\boldsymbol{\Sigma}\mathcal{D}\lambda \tag{A.34}$$

Now we will consider the integration over the $\lambda_{\alpha\beta}$ variables. Since $\mathbf{\Sigma}_{jj} > 0$ and the eigenvalues of $\mathbf{\Lambda}\mathbf{\Lambda}^\top$ are non-negative, as a matter of fact for any pair $(\lambda, \mathbf{v})$ the following holds $\lambda = \frac{\mathbf{v}^T\mathbf{\Lambda}^T\mathbf{\Lambda}\mathbf{v}}{\mathbf{v}^T\mathbf{v}} = \frac{\|\mathbf{\Lambda}\mathbf{v}\|^2}{\|\mathbf{v}\|^2} \geq 0$, the maximizer of the argument in the exponential is $\mathbf{\Lambda} = 0$, and since the argument is going to be summed over the $n \to \infty$ eigenvalues $\mathbf{\Sigma}_{jj}$, by the saddle point approximation we can consider only an expansion around $\mathbf{\Lambda} = 0$. We can then use the same analysis done in [72] and decompose the log determinant via $logdet|\mathbf{I} + \mathbf{X}| = \sum_{\xi=1} \frac{(-1)^{\xi+1}}{\xi} tr(\mathbf{X}^\xi)$.

Then it follows that

$$E_{2k}^{(n)} = \int exp\left(-\sum_{j=1}^{n} \frac{1}{2} \log det(\mathbf{I} + \frac{\sigma_x^2\sigma_w^2}{n}\mathbf{\Sigma}_{jj}\mathbf{\Lambda}\mathbf{\Lambda}^\top) - itr\mathbf{\Lambda}^\top\mathbf{Z}\right) F(z)\mathcal{D}z\mathcal{D}\mathbf{\Sigma}\mathcal{D}\lambda \tag{A.35}$$

$$= \int e^{-\sum_j \frac{\sigma_x^2\sigma_w^2}{2n}\mathbf{\Sigma}_{jj} tr(\mathbf{\Lambda}\mathbf{\Lambda}^\top) - \frac{1}{2}\sum_{j=1}^{n}\sum_{\xi\geq 2}\frac{(-1)^{\xi+1}}{\xi} tr\left(\left(\frac{\sigma_x^2\sigma_w^2}{n}\mathbf{\Sigma}_{jj}\mathbf{\Lambda}\mathbf{\Lambda}^\top\right)^\xi\right) - itr\mathbf{\Lambda}^\top\mathbf{Z}} F(z)\mathcal{D}z\mathcal{D}\mathbf{\Sigma}\mathcal{D}\lambda \tag{A.36}$$

$$= \int e^{-\frac{\sigma_x^2\sigma_w^2}{2n} tr\mathbf{\Sigma} tr(\mathbf{\Lambda}\mathbf{\Lambda}^\top)} e^{-\frac{1}{2}\sum_{\xi\geq 2}\frac{(-1)^{\xi+1}}{\xi} tr(\mathbf{\Sigma}^\xi) tr\left(\left(\frac{\sigma_x^2\sigma_w^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top\right)^\xi\right) - itr\mathbf{\Lambda}^\top\mathbf{Z}} F(z)\mathcal{D}z\mathcal{D}\mathbf{\Sigma}\mathcal{D}\lambda \tag{A.37}$$

We now consider the following change of variable

$$\bar{\lambda}_{ij} = \frac{\sigma_w\sigma_x}{\sqrt{n}}\lambda_{ij} \tag{A.38}$$

$$\mathcal{D}\bar{\mathbf{\Lambda}} = \prod_{(\alpha,\beta)\in\mathcal{Z}} \frac{d\bar{\lambda}_{\alpha\beta}}{2\pi\sigma_x\sigma_w/\sqrt{n}} \tag{A.39}$$

therefore

$$E_{2k}^{(n)} = \int e^{-\frac{\sigma_x^2\sigma_w^2}{2n} tr\mathbf{\Sigma} tr(\mathbf{\Lambda}\mathbf{\Lambda}^\top)} e^{-\frac{1}{2}\sum_{\xi\geq 2}\frac{(-1)^{\xi+1}}{\xi} tr\mathbf{\Sigma}^\xi tr\left(\left(\frac{\sigma_x^2\sigma_w^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top\right)^\xi\right) - itr\mathbf{\Lambda}^\top Z} F(z)\mathcal{D}z\mathcal{D}\mathbf{\Sigma}\mathcal{D}\lambda \tag{A.40}$$

$$= \int e^{-\frac{tr\mathbf{\Sigma}}{2} tr(\bar{\mathbf{\Lambda}}\bar{\mathbf{\Lambda}}^\top)} e^{-\frac{1}{2}\sum_{\xi\geq 2}\frac{(-1)^{\xi+1}}{\xi} tr\mathbf{\Sigma}^\xi tr\left(\left(\bar{\mathbf{\Lambda}}\bar{\mathbf{\Lambda}}^\top\right)^\xi\right) - itr\frac{\sqrt{n}}{\sigma_w\sigma_x}\bar{\mathbf{\Lambda}}^\top\mathbf{Z}} F(z)\mathcal{D}z\mathcal{D}\mathbf{\Sigma}\mathcal{D}\bar{\mathbf{\Lambda}}. \tag{A.41}$$

We will study the contribution of the exponential $e^{-\frac{1}{2}\sum_{\xi\geq 2}\frac{(-1)^{\xi+1}}{\xi}tr\mathbf{\Sigma}^{\xi}tr\left(\left(\bar{\mathbf{\Lambda}}\bar{\mathbf{\Lambda}}^{\top}\right)^{\xi}\right)}$ by considering its Taylor expansion $e^{x}=\sum_{\nu=0}^{\infty}x^{\nu}/\nu!$

$$
\begin{aligned}
E_{2k}^{(n)} &= \int\left(\sum_{\nu=0}^{\infty}\frac{\left(-\frac{1}{2}\sum_{\xi\geq 2}\frac{(-1)^{\xi+1}}{\xi}tr\mathbf{\Sigma}^{\xi}tr\left(\left(\bar{\mathbf{\Lambda}}\bar{\mathbf{\Lambda}}^{\top}\right)^{\xi}\right)\right)^{\nu}}{\nu!}\right) \\
&\qquad\cdot e^{-\frac{tr\mathbf{\Sigma}}{2}tr(\bar{\mathbf{\Lambda}}\bar{\mathbf{\Lambda}}^{\top})-itr\frac{\sqrt{n}}{\sigma_w\sigma_x}\bar{\mathbf{\Lambda}}^{\top}\mathbf{Z}}F(z)\mathcal{D}z\mathcal{D}\mathbf{\Sigma}\mathcal{D}\bar{\mathbf{\Lambda}} \qquad \text{(A.42)}\\
&= \int\left(\sum_{\nu=0}^{\infty}\frac{\left(-\frac{1}{2}\sum_{\xi\geq 2}\frac{(-1)^{\xi+1}}{\xi}tr\mathbf{\Sigma}^{\xi}tr\left(\left(\bar{\mathbf{\Lambda}}\bar{\mathbf{\Lambda}}^{\top}\right)^{\xi}\right)\right)^{\nu}}{\nu!}\right) \\
&\qquad\cdot e^{-\sum_{\lambda_{\alpha\beta}\in\mathcal{Z}}\left(\frac{tr\mathbf{\Sigma}}{2}\bar{\lambda}_{\alpha\beta}^{2}-i\ \frac{\sqrt{n}}{\sigma_w\sigma_x}\bar{\lambda}_{\alpha\beta}z_{\alpha\beta}\right)}F(z)\mathcal{D}z\mathcal{D}\mathbf{\Sigma}\mathcal{D}\bar{\mathbf{\Lambda}} \qquad \text{(A.43)}\\
&= \int\left(\sum_{\nu=0}^{\infty}\frac{\left(-\frac{1}{2}\sum_{\xi\geq 2}\frac{(-1)^{\xi+1}}{\xi}tr\mathbf{\Sigma}^{\xi}tr\left(\left(\bar{\mathbf{\Lambda}}\bar{\mathbf{\Lambda}}^{\top}\right)^{\xi}\right)\right)^{\nu}}{\nu!}\right) \\
&\qquad\cdot\left(\prod_{\lambda_{\alpha\beta}\in\mathcal{Z}}e^{-\frac{tr\mathbf{\Sigma}}{2}\bar{\lambda}_{\alpha\beta}^{2}-i\ \frac{\sqrt{n}}{\sigma_w\sigma_x}\bar{\lambda}_{\alpha\beta}z_{\alpha\beta}}\right)F(z)\mathcal{D}z\mathcal{D}\mathbf{\Sigma}\mathcal{D}\bar{\mathbf{\Lambda}} \qquad \text{(A.44)}
\end{aligned}
$$

□

A.1.2 Proof of Lemma 3.2.3 - $E_2^{(n)}$

Proof. When $k=1$, following [71] the zeroth order expansion of the Taylor series

$$\left(\sum_{\nu=0}^{\infty}\frac{\left(-\frac{1}{2}\sum_{\xi\geq 2}\frac{(-1)^{\xi+1}}{\xi}tr\mathbf{\Sigma}^{\xi}tr\left(\left(\bar{\mathbf{\Lambda}}\bar{\mathbf{\Lambda}}^{\top}\right)^{\xi}\right)\right)^{\nu}}{\nu!}\right)$$

in the integral of Lemma A.1.1 is the leading contribution to $E_2^{(n)}$.

Defining the trace of the power matrices as $\tilde{\mu}_k^{(n)}$, i.e. $\tilde{\mu}_k^{(n)}=tr(\mathbf{\Sigma}^k)/n$, the zeroth order is determined as follows.

$$
\begin{aligned}
&\int\left(\frac{\sqrt{n}}{2\pi\sigma_x\sigma_w}e^{-\frac{n\tilde{\mu}_1^{(n)}}{2}tr(\bar{\mathbf{\Lambda}}\bar{\mathbf{\Lambda}}^{\top})}e^{-itr\frac{\sqrt{n}}{\sigma_w\sigma_x}\bar{\mathbf{\Lambda}}\mathbf{Z}}F(z)\right)dz_{i_1\mu_1}d\bar{\lambda}_{i_1\mu_1}\mathcal{D}\mathbf{\Sigma} \\
&=\int\left(\int\frac{\sqrt{n}}{2\pi\sigma_x\sigma_w}e^{-\frac{n\tilde{\mu}_1^{(n)}}{2}\lambda^2-i\frac{\sqrt{n}}{\sigma_w\sigma_x}\bar{\lambda}z}\phi(z)^2dzd\bar{\lambda}\right)\mathcal{D}\mathbf{\Sigma} \qquad \text{(A.45)}
\end{aligned}
$$

$$= \int \left(\int \frac{\sqrt{n}}{2\pi\sigma_x\sigma_w} e^{-\frac{z^2}{2\tilde{\mu}_1^{(n)}\sigma_w^2\sigma_x^2}} \frac{\sqrt{2\pi}}{\sqrt{n\tilde{\mu}_1^{(n)}}} \phi(z)^2 \right.$$
$$\left. \left(\int \frac{\sqrt{n\tilde{\mu}_1^{(n)}}}{\sqrt{2\pi}} e^{-\frac{n\tilde{\mu}_1^{(n)}}{2}\left(\lambda + \frac{i\sqrt{n}}{n\tilde{\mu}_1^{(n)}\sigma_w\sigma_x} z\right)^2} d\bar{\lambda} \right) dz \right) \mathcal{D}\boldsymbol{\Sigma} \tag{A.46}$$

$$= \int \frac{\sqrt{n}}{2\pi\sigma_x\sigma_w} \phi(z)^2 e^{-\frac{z^2}{2\tilde{\mu}_1^{(n)}\sigma_w^2\sigma_x^2}} \frac{\sqrt{2\pi}}{\sqrt{n\tilde{\mu}_1^{(n)}}} dz\mathcal{D}\boldsymbol{\Sigma} \tag{A.47}$$

$$= \int \frac{1}{\sqrt{2\pi\tilde{\mu}_1^{(n)}}\sigma_x\sigma_w} \phi(z)^2 e^{-\frac{z^2}{2\tilde{\mu}_1^{(n)}\sigma_w^2\sigma_x^2}} dz\mathcal{D}\boldsymbol{\Sigma} \tag{A.48}$$

$$= \int \frac{1}{\sqrt{2\pi\tilde{\mu}_1^{(n)}}\sigma_x\sigma_w} \phi(\sqrt{\tilde{\mu}_1^{(n)}}\sigma_w\sigma_x\tilde{z})^2 e^{-\frac{\tilde{z}^2}{2}} \sqrt{\tilde{\mu}_1^{(n)}}\sigma_w\sigma_x d\tilde{z}\mathcal{D}\boldsymbol{\Sigma} \tag{A.49}$$

$$= \int \frac{1}{\sqrt{2\pi}} \phi(\sqrt{\tilde{\mu}_1^{(n)}}\sigma_w\sigma_x\tilde{z})^2 e^{-\frac{\tilde{z}^2}{2}} d\tilde{z}\mathcal{D}\boldsymbol{\Sigma} = \int \tilde{\theta}_1{}^{(n)}\mathcal{D}\boldsymbol{\Sigma} \tag{A.50}$$

□

For higher order terms of the exponential Taylor series we first notice that

$$\left(-\frac{1}{2} \sum_{\xi \geq 2} \frac{(-1)^{\xi+1}}{\xi} tr\boldsymbol{\Sigma}^\xi tr\left(\left(\bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{\Lambda}}^\top \right)^\xi \right) \right)^\nu = \left(-\frac{1}{2} \sum_{\xi \geq 2} \frac{(-1)^{\xi+1}}{\xi} (tr\boldsymbol{\Sigma}^\xi) \bar{\lambda}^{2\xi} \right)^\nu.$$

For each of the expansion terms, the same steps in (A.45)-(A.48) lead to the following $2\xi\nu$-th non-central moment of a Gaussian to be considered

$$\int \frac{\sqrt{n\tilde{\mu}_1^{(n)}}}{\sqrt{2\pi}} \bar{\lambda}^{2\xi\nu} e^{-\frac{n\tilde{\mu}_1^{(n)}}{2}(\bar{\lambda} + \frac{i\sqrt{n}z}{n\sigma_x\sigma_w\tilde{\mu}_1^{(n)}})} d\bar{\lambda} = \left(\frac{1}{n\tilde{\mu}_1^{(n)}} \right)^{\xi\nu/2} 2^{\xi\nu} \frac{\Gamma(\frac{2\xi\nu+1}{2})}{\sqrt{\pi}} \Phi(-\frac{2\xi\nu}{2}; \frac{1}{2}; -\frac{z^2}{\sigma_w^2\sigma_x^2\tilde{\mu}_1^{(n)}}) \tag{A.51}$$

where the solution of the non-central moment is given in [107] with

$$\Phi(-\frac{2\xi\nu}{2}; \frac{1}{2}; -\frac{z^2}{\sigma_w^2\sigma_x^2\tilde{\mu}_1^{(n)}}) = \sum_{i=1}^{\infty} \frac{1}{i!} \frac{(-\frac{2\xi\nu}{2})(-\frac{2\xi\nu}{2}+1)...(-\frac{2\xi\nu\nu}{2}+i-1)}{(-\frac{1}{2})(-\frac{1}{2}+1)...(-\frac{1}{2}+i-1)} \left(\frac{-z^2}{\sigma_w^2\sigma_x^2\tilde{\mu}_1^{(n)}} \right)^i. \tag{A.52}$$

This leads to the computation of the following integral

$$\int \sum_i \alpha_i \left(\frac{-z^2}{\sigma_w^2 \sigma_x^2 \tilde{\mu}_1^{(n)}} \right)^i \phi^2(\sigma_x^2 \sigma_w^2 \tilde{\mu}_1^{(n)} z) \mathcal{D}z \tag{A.53}$$

which is finite since $\Phi(-\frac{2\xi\nu}{2}; \frac{1}{2}; -\frac{z^2}{\sigma_w^2 \sigma_x^2})$ is an entire function of $\xi\nu$ and z, and when solving it by parts we compute the following integrals that are finite by hypothesis

$$\int \phi^{(k)}(\sigma_x^2 \sigma_w^2 \tilde{\mu}_1^{(n)} z) \mathcal{D}z \tag{A.54}$$

for any derivative k. However, because of the non-central moment, we also gain a factor $\left(\frac{1}{n\tilde{\mu}_1^{(n)}} \right)^{\xi\nu/2}$ and therefore these terms have a contribution that is $\mathcal{O}(\frac{1}{n})$ relative to the zeroth-order one.

Therefore the contribution of a 2-dimensional cycle is

$$E_2^{(n)} = \int \tilde{\theta}_1^{(n)} \mathcal{D}\mathbf{\Sigma} \left(1 + \mathcal{O}\left(\frac{1}{n} \right) \right). \tag{A.55}$$

A.1.3 Lemma 3.2.4 - $E_{2k}^{(n)}$

When $k > 1$, we notice that the term $\left(\prod_{\bar{\lambda}_{\alpha\beta} \in \mathcal{Z}} e^{-\frac{tr\Sigma}{2} \bar{\lambda}_{\alpha\beta}^2 - i \frac{\sqrt{n}}{\sigma_w \sigma_x} \bar{\lambda}_{\alpha\beta} z_{\alpha\beta}} \right)$ in Lemma A.1.1 is separable in variable $\bar{\lambda}_{\alpha\beta}$. If the same variable, $\bar{\lambda}_{\alpha\beta}$, does not appear in the expanded terms of the series

$$\left(\sum_{\nu=0}^{\infty} \frac{\left(-\frac{1}{2} \sum_{\xi \geq 2} \frac{(-1)^{\xi+1}}{\xi} tr\mathbf{\Sigma}^{\xi} tr\left(\left(\bar{\mathbf{\Lambda}} \bar{\mathbf{\Lambda}}^{\top} \right)^{\xi} \right) \right)^{\nu}}{\nu!} \right),$$

then we can isolate in $E_{2k}^{(n)}$ the integral $\int \phi(z) \int e^{-\frac{tr\mathbf{\Sigma}}{2} \bar{\lambda}^2 - i \frac{\sqrt{n}}{\sigma_w \sigma_x} \bar{\lambda} z} \frac{d\bar{\lambda}}{2\pi\sigma_x\sigma_w/\sqrt{n}} dz$ which is null as shown in Lemma A.1.2.

Lemma A.1.2. *Consider an odd activation function ϕ, then*

$$\int \phi(z) \int e^{-\frac{tr\mathbf{\Sigma}}{2} \bar{\lambda}^2 - i \frac{\sqrt{n}}{\sigma_w \sigma_x} \bar{\lambda} z} \frac{d\bar{\lambda}}{2\pi\sigma_x\sigma_w/\sqrt{n}} dz = 0 \tag{A.56}$$

where $\mathbf{\Sigma}$ is a positive definite matrix, and σ_x, σ_w, and n are all positive scalars.

Proof.

$$\int \phi(z) \int e^{-\frac{tr\boldsymbol{\Sigma}}{2}\bar{\lambda}^2 - i\ \frac{\sqrt{n}}{\sigma_w\sigma_x}\bar{\lambda}z} \frac{d\bar{\lambda}}{2\pi\sigma_x\sigma_w/\sqrt{n}} dz$$

$$= \int \phi(z) \int \exp\left\{ -\left(\sqrt{\frac{tr\boldsymbol{\Sigma}}{2}}\bar{\lambda} + \frac{i\sqrt{n}z}{\sigma_w\sigma_x\sqrt{2tr\boldsymbol{\Sigma}}} \right)^2 - \frac{nz^2}{\sigma_x^2\sigma_w^2 2tr\boldsymbol{\Sigma}} \right\} \frac{d\bar{\lambda}}{2\pi\sigma_x\sigma_w/\sqrt{n}} dz \tag{A.57}$$

$$= \int \frac{1}{\sqrt{2\pi}\sigma_x\sigma_w/\sqrt{n}} \phi(z) e^{-\frac{nz^2}{\sigma_x^2\sigma_w^2 2tr\boldsymbol{\Sigma}}} \int \frac{1}{\sqrt{2\pi}} e^{-\left(\sqrt{\frac{tr\boldsymbol{\Sigma}}{2}}\bar{\lambda} + \frac{i\sqrt{n}z}{\sigma_w\sigma_x\sqrt{2tr\boldsymbol{\Sigma}}}\right)^2} d\bar{\lambda} dz \tag{A.58}$$

$$= \int \frac{\sqrt{tr(\boldsymbol{\Sigma})n}}{\sqrt{2\pi\sigma_x^2\sigma_w^2}} \phi(z) e^{-\frac{nz^2}{\sigma_x^2\sigma_w^2 2tr\boldsymbol{\Sigma}}} \underbrace{\int \frac{1}{\sqrt{2\pi tr(\boldsymbol{\Sigma})}} e^{-\left(\sqrt{\frac{tr\boldsymbol{\Sigma}}{2}}\bar{\lambda} + \frac{i\sqrt{n}z}{\sigma_w\sigma_x\sqrt{2tr\boldsymbol{\Sigma}}}\right)^2} d\bar{\lambda}}_{=1} dz \tag{A.59}$$

$$= \int \frac{\sqrt{tr(\boldsymbol{\Sigma})n}}{\sqrt{2\pi\sigma_x^2\sigma_w^2}} \phi(z) e^{-\frac{nz^2}{\sigma_x^2\sigma_w^2 2tr\boldsymbol{\Sigma}}} dz = 0 \tag{A.60}$$

where the last equality is due to ϕ being an odd activation function. □

Consequently, the contribution in $E_{2k}^{(n)}$ is given by the addends in

$$\left(\sum_{\nu=0}^{\infty} \frac{\left(-\frac{1}{2} \sum_{\xi \geq 2} \frac{(-1)^{\xi+1}}{\xi} tr\boldsymbol{\Sigma}^{\xi} tr\left(\left(\bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{\Lambda}}^{\top} \right)^{\xi} \right) \right)^{\nu}}{\nu!} \right)$$

that contain all the variables $\bar{\lambda}_{\alpha\beta}$ with exponent higher or equal than one. Lemma A.1.3 allows us to consider the matrix $\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\top}$ as an adjacency matrix and allows to understand which term ξ to consider.

Lemma A.1.3. *[72, Supplementary Material, Lemma 1] Given the matrix $\boldsymbol{\Lambda}$, there exists matrix $\boldsymbol{A}$ such that*

$$tr(\boldsymbol{\Lambda}\boldsymbol{\Lambda}^{\top})^{\xi} = \frac{1}{2} tr\boldsymbol{A}^{2\xi} \tag{A.61}$$

where $\boldsymbol{A}$ is the weighted adjacency matrix defined by the undirected bigraph with vertex set $V = (I, U)$ where

$$I \equiv \{2i \mid \exists \mu s.t. (i, \mu) \in \mathcal{Z}\} \tag{A.62}$$

$$U \equiv \{2\mu - 1 \mid \exists i s.t.(i,\mu) \in \mathcal{Z}\} \tag{A.63}$$

and edges

$$E \equiv \{\{2\mu - i, 2i\} \mid (i,\mu) \in \mathcal{Z}\} \tag{A.64}$$

with weights $w(\{2\mu - i, 2i\}) = \lambda_{i\mu}$.

Consequently the trace $tr\left(\left(\bar{\mathbf{\Lambda}}\bar{\mathbf{\Lambda}}^\top\right)^\xi\right)$ corresponds to the sum over the different indices i of the product of the weights $\bar{\lambda}_{\alpha\beta}$ defining a closed $2j$ dimensional cycle starting and finishing in i, e.g. when $\xi = 1$ we will find that $tr(\bar{\mathbf{\Lambda}}\bar{\mathbf{\Lambda}}^\top) = \sum_{q=1}^{k}(\bar{\lambda}_{i_q\mu_q}\bar{\lambda}_{i_q\mu_q} + \bar{\lambda}_{i_q\mu_{q-1}}\bar{\lambda}_{i_q\mu_{q-1}})$. By construction of this proof, we are considering a unique k-dimensional cycle and this implies that only for $\xi \geq k$ there are addends in $\sum_{\xi\geq 2}\frac{(-1)^{\xi+1}}{\xi}tr\mathbf{\Sigma}^\xi tr\left(\left(\bar{\mathbf{\Lambda}}\bar{\mathbf{\Lambda}}^\top\right)^\xi\right)$ containing all the non-zero variables $\bar{\lambda}_{\alpha\beta}$ with exponent higher or equal than one. Specifically, in the next section it is going to be shown that $E_{2k}^{(n)}$ is determined to the leading order by the term for which all the non-zero variables $\bar{\lambda}_{\alpha\beta}$ appear with exponent equal to exactly one.

A.1.3.1 Proof of Lemma 3.2.4

Proof. Let's first consider the case where all the non-zero variables $\bar{\lambda}_{\alpha\beta}$ appear with exponent equal to exactly one. Since we are considering a $2k$ cycle on the i and μ indices, for $\xi = k$ and $\nu = 1$ all the variables $\lambda_{\alpha\beta}$ appear with multiplicativity $2k$ and exponent exactly equal to 1. The contribution of this term in (A.2) is

$$\begin{aligned}
&\frac{(-1)^k}{2k}\int tr\mathbf{\Sigma}^k\left(2k\prod_{(\alpha,\beta)\in\mathcal{Z}}\bar{\lambda}_{\alpha\beta}\right)\left(\prod_{(\alpha,\beta)\in\mathcal{Z}}e^{-\frac{tr\Sigma}{2}\bar{\lambda}_{\alpha\beta}^2 - i\ \frac{\sqrt{n}}{\sigma_w\sigma_x}\bar{\lambda}_{\alpha\beta}z_{\alpha\beta}}\right)F(z)\mathcal{D}\mathbf{Z}\mathcal{D}\mathbf{\Sigma}\mathcal{D}\bar{\mathbf{\Lambda}} \\
&= (-1)^k\int tr\mathbf{\Sigma}^k\left(\prod_{(\alpha,\beta)\in\mathcal{Z}}\bar{\lambda}_{\alpha\beta}e^{-\frac{tr\Sigma}{2}\bar{\lambda}_{\alpha\beta}^2 - i\ \frac{\sqrt{n}}{\sigma_w\sigma_x}\bar{\lambda}_{\alpha\beta}z_{\alpha\beta}}\phi(z_{\alpha\beta})\right)\mathcal{D}z\mathcal{D}\mathbf{\Sigma}\mathcal{D}\bar{\mathbf{\Lambda}}.
\end{aligned} \tag{A.65}$$

To simplify the notation, we now consider the new variable

$$\tilde{\mu}_k^{(n)} = \frac{1}{n}tr(\mathbf{\Sigma}^k) \tag{A.66}$$

and find that the contribution is

$$(-1)^k\int tr\mathbf{\Sigma}^k\left(\prod_{(\alpha,\beta)\in\mathcal{Z}}\bar{\lambda}_{\alpha\beta}e^{-\frac{tr\Sigma}{2}\bar{\lambda}_{\alpha\beta}^2 - i\ \frac{\sqrt{n}}{\sigma_w\sigma_x}\bar{\lambda}_{\alpha\beta}z_{\alpha\beta}}\phi(z_{\alpha\beta})\right)\mathcal{D}z\mathcal{D}\mathbf{\Sigma}\mathcal{D}\bar{\mathbf{\Lambda}}$$

$$= (-1)^k \int n\tilde{\mu}_k^{(n)} \left(\prod_{(\alpha,\beta)\in\mathcal{Z}} \bar{\lambda}_{\alpha\beta} e^{-\frac{n\tilde{\mu}_1^{(n)}}{2}\bar{\lambda}_{\alpha\beta}^2 - i\ \frac{\sqrt{n}}{\sigma_w\sigma_x}\bar{\lambda}_{\alpha\beta}z_{\alpha\beta}} \phi(z_{\alpha\beta}) \right) \left(\prod_{\bar{\lambda}_{\alpha\beta}} \frac{\sqrt{n}}{2\pi\sigma_x\sigma_w} d\bar{\lambda}_{\alpha\beta} \right) \mathcal{D}z\mathcal{D}\boldsymbol{\Sigma} \tag{A.67}$$

$$= (-1)^k n \int \tilde{\mu}_k^{(n)} \left[\int \frac{\sqrt{n}}{2\pi\sigma_x\sigma_w} e^{-\frac{n\tilde{\mu}_1^{(n)}}{2}\bar{\lambda}_{\alpha\beta}^2} e^{-i\frac{\sqrt{n}}{\sigma_w\sigma_x}\bar{\lambda}_{\alpha\beta}z_{\beta\alpha}} \phi(z_{\beta\alpha}) \bar{\lambda}_{\alpha\beta} dz_{\beta\alpha} d\bar{\lambda}_{\alpha\beta} \right]^{2k} \mathcal{D}\boldsymbol{\Sigma} \tag{A.68}$$

$$= (-1)^k n \int \tilde{\mu}_k^{(n)} \left[\int \frac{\sqrt{n}}{2\pi\sigma_x\sigma_w} e^{-\left(\sqrt{\frac{n\tilde{\mu}_1^{(n)}}{2}}\bar{\lambda} + \frac{i\sqrt{n}z}{\sigma_w\sigma_x\sqrt{2n\tilde{\mu}_1^{(n)}}}\right)^2 - \frac{nz^2}{\sigma_x^2\sigma_w^2 2n\tilde{\mu}_1^{(n)}}} \phi(z)\bar{\lambda} dz d\bar{\lambda} \right]^{2k} \mathcal{D}\boldsymbol{\Sigma} \tag{A.69}$$

$$= (-1)^k n \int \tilde{\mu}_k^{(n)} \left[\int \frac{\sqrt{n}}{2\pi\sigma_x\sigma_w} e^{-\left(\sqrt{\frac{n\tilde{\mu}_1^{(n)}}{2}}\bar{\lambda} + \frac{i\sqrt{n}z}{\sigma_w\sigma_x\sqrt{2n\tilde{\mu}_1^{(n)}}}\right)^2} e^{-\frac{nz^2}{\sigma_x^2\sigma_w^2 2n\tilde{\mu}_1^{(n)}}} \phi(z)\bar{\lambda} dz d\bar{\lambda} \right]^{2k} \mathcal{D}\boldsymbol{\Sigma} \tag{A.70}$$

$$= (-1)^k n \int \tilde{\mu}_k^{(n)} \left[\int \frac{\sqrt{n}}{2\pi\sigma_x\sigma_w} e^{-\frac{nz^2}{\sigma_x^2\sigma_w^2 2n\tilde{\mu}_1^{(n)}}} \phi(z) \left(\int e^{-\left(\sqrt{\frac{n\tilde{\mu}_1^{(n)}}{2}}\bar{\lambda} + \frac{i\sqrt{n}z}{\sigma_w\sigma_x\sqrt{2n\tilde{\mu}_1^{(n)}}}\right)^2} \bar{\lambda} d\bar{\lambda} \right) dz \right]^{2k} \mathcal{D}\boldsymbol{\Sigma} \tag{A.71}$$

$$= (-1)^k n \int \tilde{\mu}_k^{(n)} \left[\int \frac{\sqrt{n}}{2\pi\sigma_x\sigma_w} e^{-\frac{nz^2}{\sigma_x^2\sigma_w^2 2n\tilde{\mu}_1^{(n)}}} \phi(z) \frac{\sqrt{2\pi}}{\sqrt{n\tilde{\mu}_1^{(n)}}} \cdot \right.$$
$$\left. \cdot \left(\int \frac{\sqrt{n\tilde{\mu}_1^{(n)}}}{\sqrt{2\pi}} \bar{\lambda} e^{-\frac{1}{2}\left(\bar{\lambda} + \frac{i\sqrt{n}z}{\sigma_w\sigma_x n\tilde{\mu}_1^{(n)}}\right)^2 / \left(\frac{1}{\sqrt{n\tilde{\mu}_1^{(n)}}}\right)^2} d\bar{\lambda} \right) dz \right]^{2k} \mathcal{D}\boldsymbol{\Sigma} \tag{A.72}$$

$$= (-1)^k n \int \tilde{\mu}_k^{(n)} \left[\int \frac{\sqrt{n}}{2\pi\sigma_x\sigma_w} e^{-\frac{nz^2}{\sigma_x^2\sigma_w^2 2n\tilde{\mu}_1^{(n)}}} \phi(z) \frac{\sqrt{2\pi}}{\sqrt{n\tilde{\mu}_1^{(n)}}} \left(-\frac{i\sqrt{n}z}{\sigma_w\sigma_x n\tilde{\mu}_1^{(n)}} \right) dz \right]^{2k} \mathcal{D}\boldsymbol{\Sigma} \tag{A.73}$$

$$= (-1)^k i^{2k} n \int \tilde{\mu}_k^{(n)} \left[\int \frac{\sqrt{n}}{\sqrt{2\pi}\sigma_x\sigma_w\sqrt{n\tilde{\mu}_1^{(n)}}} e^{-\frac{nz^2}{\sigma_x^2\sigma_w^2 2n\tilde{\mu}_1^{(n)}}} \phi(z) \left(\frac{\sqrt{n}z}{\sigma_w\sigma_x n\tilde{\mu}_1^{(n)}} \right) dz \right]^{2k} \mathcal{D}\boldsymbol{\Sigma} \tag{A.74}$$

$$= n \int \tilde{\mu}_k^{(n)} \left[\int \frac{nz}{\sqrt{2\pi}\sigma_x^2\sigma_w^2(n\tilde{\mu}_1^{(n)})^{3/2}} e^{-\frac{nz^2}{\sigma_x^2\sigma_w^2 2n\tilde{\mu}_1^{(n)}}} \phi(z)dz \right]^{2k} \mathcal{D}\boldsymbol{\Sigma} \tag{A.75}$$

We now consider the following change of variable

$$\tilde{z} = \frac{\sqrt{n}}{\sigma_w\sigma_x\sqrt{n\tilde{\mu}_1^{(n)}}} z \tag{A.76}$$

then

$$n \int \tilde{\mu}_k^{(n)} \left[\int \frac{nz}{\sqrt{2\pi}\sigma_x^2\sigma_w^2(n\tilde{\mu}_1^{(n)})^{3/2}} e^{-\frac{nz^2}{\sigma_x^2\sigma_w^2 2n\tilde{\mu}_1^{(n)}}} \phi(z)dz \right]^{2k} \mathcal{D}\boldsymbol{\Sigma}$$

$$= n \int \tilde{\mu}_k^{(n)} \left[\int \frac{n}{\sqrt{2\pi}\sigma_x^2\sigma_w^2(n\tilde{\mu}_1^{(n)})^{3/2}} \frac{\sigma_w^2\sigma_x^2 n\tilde{\mu}_1^{(n)}}{n} \tilde{z} e^{-\frac{\tilde{z}^2}{2}} \phi(\frac{\sigma_w\sigma_x\sqrt{n\tilde{\mu}_1^{(n)}}}{\sqrt{n}}\tilde{z})d\tilde{z} \right]^{2k} \mathcal{D}\boldsymbol{\Sigma} \tag{A.77}$$

$$= n \int \tilde{\mu}_k^{(n)} \left[\int \frac{1}{\sqrt{2\pi n\tilde{\mu}_1^{(n)}}} \tilde{z} e^{-\frac{\tilde{z}^2}{2}} \phi(\frac{\sigma_w\sigma_x\sqrt{n\tilde{\mu}_1^{(n)}}}{\sqrt{n}}\tilde{z})d\tilde{z} \right]^{2k} \mathcal{D}\boldsymbol{\Sigma} \tag{A.78}$$

$$= n \int \tilde{\mu}_k^{(n)} \left[\int \frac{1}{\sqrt{2\pi n\tilde{\mu}_1^{(n)}}} \frac{\sigma_w\sigma_x\sqrt{n\tilde{\mu}_1^{(n)}}}{\sqrt{n}} e^{-\frac{\tilde{z}^2}{2}} \phi'(\frac{\sigma_w\sigma_x\sqrt{n\tilde{\mu}_1^{(n)}}}{\sqrt{n}}\tilde{z})d\tilde{z} \right]^{2k} \mathcal{D}\boldsymbol{\Sigma} \tag{A.79}$$

$$= n^{1-k} \int \tilde{\mu}_k^{(n)} \left[\int \frac{\sigma_w\sigma_x}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \phi'(\sigma_w\sigma_x\sqrt{\tilde{\mu}_1^{(n)}}\tilde{z})d\tilde{z} \right]^{2k} \mathcal{D}\boldsymbol{\Sigma} = n^{1-k} \int \tilde{\theta}_2^{(n)k} \tilde{\mu}_k^{(n)} \mathcal{D}\boldsymbol{\Sigma} \tag{A.80}$$

where the second to last equation is due to integration by parts and we consider

$$\tilde{\theta}_2^{(n)} = \left(\int \frac{\sigma_w\sigma_x}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \phi'(\sigma_w\sigma_x\sqrt{\tilde{\mu}_1^{(n)}}\tilde{z})d\tilde{z} \right)^2.$$

Let's now consider any case where some variables $\bar{\lambda}_{\alpha\beta}$ appear with exponent higher than one. Following the same steps as in (A.65)-(A.72), it is possible to decompose the contribution into the product of integrals uniquely dependent on the edge $(\alpha, \beta) \in \mathcal{Z}$. Specifically, each of these integrals contain the computations of the moment of $\bar{\lambda}_{\alpha\beta}$ of the same order as their exponent in the expansion of the

series. From [107], the moments are

$$\mathbb{E}[\lambda^\kappa] = \begin{cases} \left(\frac{1}{n\tilde{\mu}_1^{(n)}}\right)^{\kappa/2} 2^{\kappa/2}\frac{\Gamma(\frac{\kappa+1}{2})}{\sqrt{\pi}}\Phi(-\frac{\kappa}{2};\frac{1}{2};-\frac{z^2}{\sigma_w^2\sigma_x^2}) & \text{if } \kappa \text{ even} \\ \left(-\frac{i\sqrt{n}z}{\sigma_w\sigma_x n\tilde{\mu}_1^{(n)}}\right)\left(\frac{1}{n\tilde{\mu}_1^{(n)}}\right)^{\frac{\kappa-1}{2}} 2^{\frac{\kappa+1}{2}}\frac{\Gamma(\frac{\kappa+1}{2})}{\sqrt{\pi}}\Phi(-\frac{1-\kappa}{2};\frac{3}{2};-\frac{z^2}{\sigma_w^2\sigma_x^2}) & \text{if } \kappa \text{ odd} \end{cases} \tag{A.81}$$

As for $k = 1$, these solutions lead to the computation of the integration of z^κ multiplied by $\phi(\sigma_x\sigma_w\sqrt{\tilde{\mu}_1^n}z)$ on the Gaussian measure, whose solution is found by using the integration by parts and is finite because of the hypothesis on the activation function ϕ. However, because of the κ moment of $\bar{\lambda}$ we gain a factor $1/n^\kappa$ and, since $\kappa > 1$, the contribution for terms in (A.2) with exponential larger than one have a contribution $\mathcal{O}(1/n)$ relative to those computed in (A.80) when all the exponents are exactly equal to 1. Furthermore, the multiplicativity of these paths is $2kk^\kappa$.

Therefore, the contribution for $k > 1$ dimensional graphs is

$$E_{2k}^{(n)} = n^{1-k}\int \tilde{\theta}_2^{(n)k}\tilde{\mu}_k^{(n)}\mathcal{D}\boldsymbol{\Sigma}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.82}$$

□

A.1.4 Proposition 3.2.5

To prove Proposition 3.2.5, it is required considering how many graphs, $\#_1$ and $\#_2$ can be generated by varying respectively the i- and the μ- vertices for each possible pattern. Since, a value among the n_1 available without reinsertion has to be assigned to each of the $k - I_i$ i-vertices

$$\#_1 = \binom{n_1}{k-I_i} = \left(\frac{n_1 e}{k-I_i}\right)^{k-I_1}(2\pi(k-I_i))^{-1/2}e^{-\frac{(k-I_1)^2}{2n_1}(1+o(1))} \tag{A.83}$$

$$= \left(\frac{n_1 e}{k-I_i}\right)^{k-I_1}(2\pi(k-I_i))^{-1/2}(1+o(\frac{1}{n_1})) \tag{A.84}$$

and equivalently

$$\#_2 = \binom{m}{k-I_\mu} = \left(\frac{me}{k-I_\mu}\right)^{k-I_\mu}(2\pi(k-I_\mu))^{-1/2}(1+o(\frac{1}{m})). \tag{A.85}$$

where the asymptotic behaviour for n_1 or m being much larger than $k-I_i$ or $k-I_\mu$ is defined in [96][1]

Further, it is also necessary to consider the expected contribution of each graph.

Proposition A.1.4. *For $\boldsymbol{M}=\frac{1}{m}\boldsymbol{Y}\boldsymbol{Y}^\top$ and $\boldsymbol{Y}=\phi\left(\boldsymbol{W}\tilde{\boldsymbol{X}}\right)$ as defined in Theorem 3.2.2, an admissible graph G with $2k$ edges, b blocks of size 1, $c=1+I_i+I_\mu-b$ cyclic blocks of dimensions $\{\tilde{k}_1,\ldots,\tilde{k}_c\}$ is such that the contribution of the graph, $E_G^{(n)}$, grows as $n_0^{1+I_i+I_\mu-k}\tilde{\theta}_2^{(n)^{k-b}}\tilde{\theta}_1^{(n)^b}\left(\prod_{c=1}^{1+I_i+I_\mu-b}\tilde{\mu}_{\tilde{k}_c}^{(n)}\right)\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)$*

Proof. When considering the contribution of a composite graph, here we will consider as an example the junction of cycles of dimensions $2k_1$ and $2k_2$ on a vertex, the contribution of the graph, $E_G^{(n)}$, can be determined by initially repeating the same steps as in Lemma A.1.1. In the Taylor expansion of the exponential,

$$\left(\sum_{\nu=0}^{\infty}\frac{\left(-\frac{1}{2}\sum_{\xi\geq 2}\frac{(-1)^{\xi+1}}{\xi}tr\boldsymbol{\Sigma}^\xi tr\left(\left(\bar{\boldsymbol{\Lambda}}\bar{\boldsymbol{\Lambda}}^\top\right)^\xi\right)\right)^\nu}{\nu!}\right),$$

as for normal cycles the leading contribution up to a relative factor $\mathcal{O}(1/n)$ is given by the terms with all of the vertices appearing with an exponential exactly equal to one. There are two cases for which the exponential is exactly on for all of the vertices: for $\nu=1$ when $\xi=k_1+k_2$ while for $\nu=2$ when the product of the terms $\xi=k_1$ and $\xi=k_2$ is considered. However, since the term $tr(\boldsymbol{\Sigma}^k)$ is $\mathcal{O}(n)$, the case for $\nu=1$ has a contribution that is $\mathcal{O}(1/n)$ compared to the case for $\nu=2$. Consequently, the contribution $E_G^{(n)}$ is determined up to a term $\mathcal{O}(1/n)$ by the inner product of the cases $\xi=k_1$ and $\xi=k_2$ when $\nu=1$. However, the total contribution $E_G^{(n)}$ does not correspond to exactly the multiplication of the cycles contributions $E_{2k_1}^{(n)}$ and $E_{2k_2}^{(n)}$ since the integration over $\mathcal{D}\boldsymbol{\Sigma}$ is not split among cycles.

The same analysis can be extended to any admissible graph and the contribution is always determined to a leading order by an integral which is separable in the product of the contributions of each single cycle.

[1]Note that for $k-I=0$, the indeterminate form $(1/0)^0/0$ is given the correct value of one, i.e. $\binom{n}{0}=1$

Since for each $(\overrightarrow{k}, I_i, I_\mu, b)$ pattern there are $n_c = 1 + I_i + I_\mu - b$ cycles whose respective dimension is $2\tilde{k}_c$ with $\sum_{c=1}^{n_c} \tilde{k}_c = k - b$, then the contribution is

$$E^{(n)}_{G(k,I_i,I_\mu,b)} = \int \left(\prod_{c=1}^{n_c} n_0^{1-\tilde{k}_c} \tilde{\theta}_2^{(n)^{\tilde{k}_c}} \tilde{\mu}^{(n)}_{\tilde{k}_c} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \right) \tilde{\theta}_1^{(n)^b} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \mathcal{D}\mathbf{\Sigma} \tag{A.86}$$

$$= n_0^{n_c} \int n_0^{-\sum_c^{n_c} \tilde{k}_c} \tilde{\theta}_2^{(n)^{\sum_c^{n_c} \tilde{k}_c}} \left(\prod_{c=1}^{n_c} \tilde{\mu}^{(n)}_{\tilde{k}_c} \right) \tilde{\theta}_1^{(n)^b} \mathcal{D}\mathbf{\Sigma} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.87}$$

$$= n_0^{1+I_i+I_\mu-b} n_0^{-k+b} \int \tilde{\theta}_2^{(n)^{k-b}} \left(\prod_{c=1}^{n_c} \tilde{\mu}^{(n)}_{\tilde{k}_c} \right) \tilde{\theta}_1^{(n)^b} \mathcal{D}\mathbf{\Sigma} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.88}$$

$$= n_0^{1+I_i+I_\mu-k} \int \tilde{\theta}_2^{(n)^{k-b}} \tilde{\theta}_1^{(n)^b} \left(\prod_{c=1}^{1+I_i+I_\mu-b} \tilde{\mu}^{(n)}_{\tilde{k}_c} \right) \mathcal{D}\mathbf{\Sigma} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right). \tag{A.89}$$

□

A.1.4.1 Proof of Proposition 3.2.5

Proof. Let's start by considering the momentum for finite n

$$\mathfrak{m}_k^{(n)} = \frac{1}{n_1 m^k} \sum_{I_i, I_\mu = 1}^{k} \sum_{b=0}^{I_i+I_\mu+1} \sum_{\overrightarrow{k}} \mathcal{B}(\overrightarrow{k}, I_i, I_\mu, b) E^{(n)}_{G(k,I_i,I_\mu,b)} \#_1 \#_2 \tag{A.90}$$

$$\begin{aligned} &= \frac{1}{n_1 m^k} \sum_{I_i, I_\mu = 1}^{k} \sum_{b=0}^{I_i+I_\mu+1} \sum_{\overrightarrow{k}} \mathcal{B}(\overrightarrow{k}, I_i, I_\mu, b) n_0^{1+I_i+I_\mu-k} \int \tilde{\theta}_2^{(n)^{k-b}} \tilde{\theta}_1^{(n)^b} \left(\prod_{c=1}^{1+I_i+I_\mu-b} \tilde{\mu}^{(n)}_{\tilde{k}_c} \right) \mathcal{D}\mathbf{\Sigma} \\ &\quad \cdot \left(\frac{n_1 e}{k - I_i}\right)^{k-I_1} (2\pi(k-I_i))^{-1/2} \left(\frac{me}{k-I_\mu}\right)^{k-I_\mu} (2\pi(k-I_\mu))^{-1/2} \\ &\quad \cdot \left(1 + o\left(\frac{1}{m}\right)\right) \left(1 + o\left(\frac{1}{n_1}\right)\right) \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \end{aligned} \tag{A.91}$$

$$\begin{aligned} &= \frac{1}{n_1 m^k} \sum_{I_i, I_\mu = 1}^{k} \sum_{b=0}^{I_i+I_\mu+1} n_0^{1+I_i+I_\mu-k} \int \tilde{\theta}_2^{(n)^{k-b}} \tilde{\theta}_1^{(n)^b} n_1^{k-I_1} m^{k-I_\mu} \\ &\quad \cdot \left(\sum_{\overrightarrow{k}} \mathcal{B}(\overrightarrow{k}, I_i, I_\mu, b) \left(\prod_{c=1}^{1+I_i+I_\mu-b} \tilde{\mu}^{(n)}_{\tilde{k}_c} \right) \right) \mathcal{D}\mathbf{\Sigma} (2\pi)^{-1} ((k-I_i)(k-I_\mu))^{-1/2} \end{aligned}$$

$$\cdot \left(\frac{e}{k-I_i}\right)^{k-I_1} \left(\frac{e}{k-I_\mu}\right)^{k-I_\mu} \left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.92}$$

$$= \sum_{I_i,I_\mu=1}^{k} \sum_{b=0}^{I_i+I_\mu+1} \int \tilde{\theta}_1^{(n)^b} \tilde{\theta}_2^{(n)^{k-b}} \frac{n_1^{k-1}}{n_0^{k-1}} \frac{n_0^{I_1}}{n_1^{I_1}} \frac{n_0^{I_\mu}}{m^{I_\mu}} \left(\sum_{\overrightarrow{k}} \mathcal{B}(\overrightarrow{k}, I_i, I_\mu, b) \left(\prod_{c=1}^{1+I_i+I_\mu-b} \tilde{\mu}_{\tilde{k}_c}^{(n)}\right)\right) \mathcal{D}\boldsymbol{\Sigma}$$
$$\cdot (2\pi)^{-1}((k-I_i)(k-I_\mu))^{-1/2} \left(\frac{e}{k-I_i}\right)^{k-I_1} \left(\frac{e}{k-I_\mu}\right)^{k-I_\mu} \left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.93}$$

$$= \varphi^{1-k} \sum_{I_i,I_\mu=1}^{k} \sum_{b=0}^{I_i+I_\mu+1} \int \frac{\tilde{\theta}_1^{(n)^b} \tilde{\theta}_2^{(n)^{k-b}} \varphi^{I_1} \psi^{I_\mu}}{2\pi((k-I_i)(k-I_\mu))^{\frac{1}{2}}} \left(\frac{e}{k-I_i}\right)^{k-I_1} \left(\frac{e}{k-I_\mu}\right)^{k-I_\mu}$$
$$\cdot \left(\sum_{\overrightarrow{k}} \mathcal{B}(\overrightarrow{k}, I_i, I_\mu, b) \left(\prod_{c=1}^{1+I_i+I_\mu-b} \tilde{\mu}_{\tilde{k}_c}^{(n)}\right)\right) \mathcal{D}\boldsymbol{\Sigma} \left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.94}$$

where we used the coefficients $\varphi = n_0/n_1$, $\psi = n_0/m$.

When considering the limiting moment for $n \to \infty$ we can use the property that almost surely $\tilde{\mu}_k^{(n)} \overset{a.s.}{\to} \mathbb{E}_{\boldsymbol{\Sigma}}\left[tr(\boldsymbol{\Sigma}^k)/n\right] = \mathbb{E}_{\rho_{\boldsymbol{\Sigma}}}\left[\lambda^k\right] =: \tilde{\mu}_k$. As a consequence, $\tilde{\mu}_k^{(n)}$ converges in distribution to $\tilde{\mu}_k$ [44, Theorem 18.5], and therefore by [44, Theorem 18.1] the following holds

$$\lim_{n\to\infty} \mathbb{E}\left[f(\tilde{\mu}_k^{(n)})\right] = \mathbb{E}\left[f(\tilde{\mu}_k)\right]. \tag{A.95}$$

for any continuous and bounded function f. In our case the integrand function is bounded thanks to the constraint on the activation function and to the eigenspectrum of $\boldsymbol{\Sigma}$ being bounded.

By then introducing the new variable

$$\mathcal{A}(k, I_i, I_\mu, b) = \sum_{\overrightarrow{k}} \mathcal{B}(\overrightarrow{k}, I_i, I_\mu, b) \left(\prod_{c=1}^{1+I_i+I_\mu-b} \tilde{\mu}_{\tilde{k}_c}\right). \tag{A.96}$$

and by defining the limit variables

$$\tilde{\theta}_1 = \int \frac{\sigma_w \sigma_x}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \phi^2(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1}\tilde{z}) d\tilde{z} \tag{A.97}$$

$$\tilde{\theta}_2 = \left(\int \frac{\sigma_w \sigma_x}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \phi'(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1}\tilde{z}) d\tilde{z}\right)^2 \tag{A.98}$$

it is straightforward to compute the limiting moment with $n \to \infty$

$$\mathfrak{m}_k = \lim_{n\to\infty} \mathfrak{m}_k^{(n)} \tag{A.99}$$

$$= \varphi^{1-k} \sum_{I_i,I_\mu=1}^{k} \sum_{b=0}^{I_i+I_\mu+1} \frac{\mathcal{A}(k,I_i,I_\mu,b)\tilde{\theta}_1^b\tilde{\theta}_2^{k-b}\varphi^{I_1}\psi^{I_\mu}}{2\pi((k-I_i)(k-I_\mu))^{\frac{1}{2}}} \left(\frac{e}{k-I_i}\right)^{k-I_1} \left(\frac{e}{k-I_\mu}\right)^{k-I_\mu} \tag{A.100}$$

$$= \varphi^{1-k} H_k. \tag{A.101}$$

We now study the rate at which the empirical moment converges to its expectation by considering their absolute difference

$$\begin{aligned}
|\mathfrak{m}_k^{(n)} - \mathfrak{m}_k| = \Bigg| & \varphi^{1-k} \sum_{I_i,I_\mu=1}^{k} \sum_{b=0}^{I_i+I_\mu+1} \Bigg[\int \frac{\tilde{\theta}_1^{(n)^b}\tilde{\theta}_2^{(n)^{k-b}}\varphi^{I_1}\psi^{I_\mu}}{2\pi((k-I_i)(k-I_\mu))^{\frac{1}{2}}} \left(\frac{e}{k-I_i}\right)^{k-I_1} \left(\frac{e}{k-I_\mu}\right)^{k-I_\mu} \\
& \cdot \left(\sum_{\overrightarrow{k}} \mathcal{B}(\overrightarrow{k}, I_i, I_\mu, b) \left(\prod_{c=1}^{1+I_i+I_\mu-b} \tilde{\mu}_{\tilde{k}_c}^{(n)} \right) \right) \mathcal{D}\boldsymbol{\Sigma} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \\
& - \frac{\mathcal{A}(k,I_i,I_\mu,b)\tilde{\theta}_1^b\tilde{\theta}_2^{k-b}\varphi^{I_1}\psi^{I_\mu}}{2\pi((k-I_i)(k-I_\mu))^{\frac{1}{2}}} \left(\frac{e}{k-I_i}\right)^{k-I_1} \left(\frac{e}{k-I_\mu}\right)^{k-I_\mu} \Bigg] \Bigg|
\end{aligned} \tag{A.102}$$

$$\begin{aligned}
\leq \varphi^{1-k} & \sum_{I_i,I_\mu=1}^{k} \sum_{b=0}^{I_i+I_\mu+1} \Bigg| \int \frac{\varphi^{I_1}\psi^{I_\mu}}{2\pi((k-I_i)(k-I_\mu))^{\frac{1}{2}}} \left(\frac{e}{k-I_i}\right)^{k-I_1} \left(\frac{e}{k-I_\mu}\right)^{k-I_\mu} \\
& \cdot \Bigg[\tilde{\theta}_1^{(n)^b}\tilde{\theta}_2^{(n)^{k-b}} \left(\sum_{\overrightarrow{k}} \mathcal{B}(\overrightarrow{k}, I_i, I_\mu, b) \left(\prod_{c=1}^{1+I_i+I_\mu-b} \tilde{\mu}_{\tilde{k}_c}^{(n)} \right) \right) \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \\
& - \mathcal{A}(k,I_i,I_\mu,b)\tilde{\theta}_1^b\tilde{\theta}_2^{k-b} \Bigg] \mathcal{D}\boldsymbol{\Sigma} \Bigg|
\end{aligned} \tag{A.103}$$

$$\begin{aligned}
= \varphi^{1-k} & \sum_{I_i,I_\mu=1}^{k} \sum_{b=0}^{I_i+I_\mu+1} \Bigg| \int \frac{\varphi^{I_1}\psi^{I_\mu}}{2\pi((k-I_i)(k-I_\mu))^{\frac{1}{2}}} \left(\frac{e}{k-I_i}\right)^{k-I_1} \left(\frac{e}{k-I_\mu}\right)^{k-I_\mu} \\
& \cdot \Bigg[\tilde{\theta}_1^{(n)^b}\tilde{\theta}_2^{(n)^{k-b}} \left(\sum_{\overrightarrow{k}} \mathcal{B}(\overrightarrow{k}, I_i, I_\mu, b) \left(\prod_{c=1}^{1+I_i+I_\mu-b} \tilde{\mu}_{\tilde{k}_c}^{(n)} \right) \right) \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \\
& - \tilde{\theta}_1^{(n)^b}\tilde{\theta}_2^{(n)^{k-b}} \mathcal{A}(k,I_i,I_\mu,b) \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \\
& + \tilde{\theta}_1^{(n)^b}\tilde{\theta}_2^{(n)^{k-b}} \mathcal{A}(k,I_i,I_\mu,b) \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right)
\end{aligned}$$

$$
\begin{aligned}
&- \tilde{\theta}_1^{(n)^b}\tilde{\theta}_2^{k-b}\mathcal{A}(k,I_i,I_\mu,b)\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)\\
&+ \tilde{\theta}_1^{(n)^b}\tilde{\theta}_2^{k-b}\mathcal{A}(k,I_i,I_\mu,b)\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)\\
&-\mathcal{A}(k,I_i,I_\mu,b)\tilde{\theta}_1^b\tilde{\theta}_2^{k-b}\Big]\,\mathcal{D}\boldsymbol{\Sigma}\Big| \qquad \text{(A.104)}\\
\leq\ & \varphi^{1-k}\sum_{I_i,I_\mu=1}^{k}\sum_{b=0}^{I_i+I_\mu+1}\frac{\varphi^{I_1}\psi^{I_\mu}}{2\pi((k-I_i)(k-I_\mu))^{\frac{1}{2}}}\left(\frac{e}{k-I_i}\right)^{k-I_1}\left(\frac{e}{k-I_\mu}\right)^{k-I_\mu}\\
&\cdot\Bigg[\Bigg|\int \tilde{\theta}_1^{(n)^b}\tilde{\theta}_2^{(n)^{k-b}}\left(\sum_{\overrightarrow{k}}\mathcal{B}(\overrightarrow{k},I_i,I_\mu,b)\left(\prod_{c=1}^{1+I_i+I_\mu-b}\tilde{\mu}_{\tilde{k}_c}^{(n)}\right)\right)\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)\\
&- \tilde{\theta}_1^{(n)^b}\tilde{\theta}_2^{(n)^{k-b}}\mathcal{A}(k,I_i,I_\mu,b)\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)\mathcal{D}\boldsymbol{\Sigma}\Bigg|\\
&+\Bigg|\int \tilde{\theta}_1^{(n)^b}\tilde{\theta}_2^{(n)^{k-b}}\mathcal{A}(k,I_i,I_\mu,b)\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)\\
&- \tilde{\theta}_1^{(n)^b}\tilde{\theta}_2^{k-b}\mathcal{A}(k,I_i,I_\mu,b)\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)\mathcal{D}\boldsymbol{\Sigma}\Bigg|\\
&+\Bigg|\int \tilde{\theta}_1^{(n)^b}\tilde{\theta}_2^{k-b}\mathcal{A}(k,I_i,I_\mu,b)\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)\\
&-\mathcal{A}(k,I_i,I_\mu,b)\tilde{\theta}_1^b\tilde{\theta}_2^{k-b}\mathcal{D}\boldsymbol{\Sigma}\Bigg|\Bigg] \qquad \text{(A.105)}\\
\leq\ & \varphi^{1-k}\sum_{I_i,I_\mu=1}^{k}\sum_{b=0}^{I_i+I_\mu+1}\frac{\varphi^{I_1}\psi^{I_\mu}}{2\pi((k-I_i)(k-I_\mu))^{\frac{1}{2}}}\left(\frac{e}{k-I_i}\right)^{k-I_1}\left(\frac{e}{k-I_\mu}\right)^{k-I_\mu}\\
&\cdot\Bigg[\Bigg|\int \tilde{\theta}_1^{(n)^b}\tilde{\theta}_2^{(n)^{k-b}}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)\\
&\cdot\left(\sum_{\overrightarrow{k}}\mathcal{B}(\overrightarrow{k},I_i,I_\mu,b)\left(\prod_{c=1}^{1+I_i+I_\mu-b}\tilde{\mu}_{\tilde{k}_c}^{(n)}\right)-\mathcal{A}(k,I_i,I_\mu,b)\right)\mathcal{D}\boldsymbol{\Sigma}\Bigg|\\
&+\Bigg|\int \tilde{\theta}_1^{(n)^b}\mathcal{A}(k,I_i,I_\mu,b)\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)\left(\tilde{\theta}_2^{(n)^{k-b}}-\tilde{\theta}_2^{k-b}\right)\mathcal{D}\boldsymbol{\Sigma}\Bigg|\\
&+\Bigg|\int \tilde{\theta}_2^{k-b}\mathcal{A}(k,I_i,I_\mu,b)\left(\tilde{\theta}_1^{(n)^b}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)-\tilde{\theta}_1^b\right)\mathcal{D}\boldsymbol{\Sigma}\Bigg|\Bigg] \qquad \text{(A.106)}
\end{aligned}
$$

We now consider all of the three absolute differences in (A.106) independently and find their upper bound relying on the fact that there are two constants K_1 and K_2 such that $\max(\tilde{\theta}_1,\tilde{\theta}_1^{(n)})\leq K_1$ and $\max(\tilde{\theta}_2,\tilde{\theta}_2^{(n)})\leq K_2$ because of the bounded

eigen-distribution of $\mathbf{\Sigma}$ and the conditions on the activation function ϕ.

The first absolute difference also relies on the fact that because of the boundedness of the eigen-spectrum of $\mathbf{\Sigma}$ there is a constant $K_{\mu_k} \geq \max(\tilde{\mu}_k^{(n)}, \tilde{\mu}_k)$. Therefore,

$$\left| \int \tilde{\theta}_1^{(n)^b} \tilde{\theta}_2^{(n)^{k-b}} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \left(\sum_{\overrightarrow{k}} \mathcal{B}(\overrightarrow{k}, I_i, I_\mu, b) \left(\prod_{c=1}^{1+I_i+I_\mu-b} \tilde{\mu}_{\tilde{k}_c}^{(n)} \right) - \mathcal{A}(k, I_i, I_\mu, b) \right) \mathcal{D}\mathbf{\Sigma} \right|$$

$$\leq K_1^b K_2^{k-b} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \sum_{\overrightarrow{k}} \mathcal{B}(\overrightarrow{k}, I_i, I_\mu, b) \int \prod_{c=1}^{1+I_i+I_\mu-b} \left| \tilde{\mu}_{\tilde{k}_c}^{(n)} - \tilde{\mu}_{\tilde{k}_c} \right| \mathcal{D}\mathbf{\Sigma} \quad \text{(A.107)}$$

$$\leq K_1^b K_2^{k-b} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \sum_{\overrightarrow{k}} \mathcal{B}(\overrightarrow{k}, I_i, I_\mu, b) \left(\prod_{c=2}^{1+I_i+I_\mu-b} K_{\mu_{\tilde{k}_c}} \right) \int \left| \tilde{\mu}_{\tilde{k}_1}^{(n)} - \tilde{\mu}_{\tilde{k}_1} \right| \mathcal{D}\mathbf{\Sigma} \quad \text{(A.108)}$$

$$= K_1^b K_2^{k-b} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \sum_{\overrightarrow{k}} \mathcal{B}(\overrightarrow{k}, I_i, I_\mu, b) \left(\prod_{c=2}^{1+I_i+I_\mu-b} K_{\mu_{\tilde{k}_c}} \right) \sqrt{\frac{2}{n\pi}} \sigma_{\mathbf{\Sigma}} \quad \text{(A.109)}$$

where in the last equality we have used the fact that the eigenvalues of $\mathbf{\Sigma}$ have variance $\sigma_{\mathbf{\Sigma}}^2$ and for large n then $\tilde{\mu}_{\tilde{k}_1}^{(n)} \sim \mathcal{N}(\mu_{\tilde{k}_1}, \frac{\sigma_{\mathbf{\Sigma}}^2}{n})$. Specifically for $X \sim \mathcal{N}(0, \sigma^2/n)$ the following also holds

$$\mathbb{E}\left[|X|\right] = \int_{-\infty}^{\infty} |x| \frac{1}{\sqrt{2\pi\sigma^2/n}} e^{-\frac{nx^2}{2\sigma^2}} dx = 2 \int_0^{\infty} x \frac{1}{\sqrt{2\pi\sigma^2/n}} e^{-\frac{nx^2}{2\sigma^2}} dx \quad \text{(A.110)}$$

$$= \sqrt{\frac{2}{\pi n}} \sigma \left[-e^{-\frac{nx^2}{2\sigma^2}} \right|_0^{\infty} = \sqrt{\frac{2}{\pi n}} \sigma. \quad \text{(A.111)}$$

The second absolute difference relies on the property that the activation function is such that there is a constant $K_{\phi'}$ such that $K_{\phi'} \geq \left| \int \frac{\sigma_w \sigma_x}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \phi'(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1} \tilde{z}) d\tilde{z} \right|$. Therefore

$$\left| \int \tilde{\theta}_1^{(n)^b} \mathcal{A}(k, I_i, I_\mu, b) \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \left(\tilde{\theta}_2^{(n)^{k-b}} - \tilde{\theta}_2^{k-b} \right) \mathcal{D}\mathbf{\Sigma} \right|$$

$$\leq K_1^b K_2^{k-b-1} \mathcal{A}(k, I_i, I_\mu, b) \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \left| \int \tilde{\theta}_2^{(n)} - \tilde{\theta}_2 \mathcal{D}\mathbf{\Sigma} \right| \quad \text{(A.112)}$$

$$= K_1^b K_2^{k-b-1} \mathcal{A}(k, I_i, I_\mu, b) \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right)$$

$$\cdot \left| \int \left(\int \frac{\sigma_w \sigma_x}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \phi'(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1^{(n)}} \tilde{z}) d\tilde{z} \right)^2 - \left(\int \frac{\sigma_w \sigma_x}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \phi'(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1} \tilde{z}) d\tilde{z} \right)^2 \mathcal{D}\boldsymbol{\Sigma} \right| \tag{A.113}$$

$$= K_1^b K_2^{k-b-1} \mathcal{A}(k, I_i, I_\mu, b) \left(1 + \mathcal{O}\left(\frac{1}{n} \right) \right)$$
$$\cdot \left| \int \left(\int \frac{\sigma_w \sigma_x}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \phi'(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1^{(n)}} \tilde{z}) d\tilde{z} - \int \frac{\sigma_w \sigma_x}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \phi'(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1} \tilde{z}) d\tilde{z} \right) \right.$$
$$\left. \cdot \left(\int \frac{\sigma_w \sigma_x}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \phi'(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1^{(n)}} \tilde{z}) d\tilde{z} + \int \frac{\sigma_w \sigma_x}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \phi'(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1} \tilde{z}) d\tilde{z} \right) \mathcal{D}\boldsymbol{\Sigma} \right| \tag{A.114}$$

$$\leq 2 K_1^b K_2^{k-b-1} K_{\phi'} \mathcal{A}(k, I_i, I_\mu, b) \left(1 + \mathcal{O}\left(\frac{1}{n} \right) \right)$$
$$\cdot \left| \int \int \frac{\sigma_w \sigma_x}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \left(\phi'(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1^{(n)}} \tilde{z}) - \phi'(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1} \tilde{z}) \right) d\tilde{z} \mathcal{D}\boldsymbol{\Sigma} \right| \tag{A.115}$$

$$= 2 K_1^b K_2^{k-b-1} K_{\phi'} \mathcal{A}(k, I_i, I_\mu, b) \left(1 + \mathcal{O}\left(\frac{1}{n} \right) \right)$$
$$\cdot \left| \int \frac{\sigma_w \sigma_x}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \int \left(\phi'(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1^{(n)}} \tilde{z}) - \phi'(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1} \tilde{z}) \right) \mathcal{D}\boldsymbol{\Sigma} d\tilde{z} \right| \tag{A.116}$$

$$= 2 K_1^b K_2^{k-b-1} K_{\phi'} \mathcal{A}(k, I_i, I_\mu, b) \left(1 + \mathcal{O}\left(\frac{1}{n} \right) \right)$$
$$\cdot \left| \int \frac{\sigma_w \sigma_x}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \int \left(\phi'(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1} \tilde{z}) - \phi'(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1} \tilde{z}) \right. \right.$$
$$\left. \left. + \sum_{q=1} \left(\phi^{(q+1)}(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1} \tilde{z}) \frac{(\sigma_w \sigma_x \tilde{z})}{2\sqrt{\tilde{\mu}_1}} \right) \frac{(\tilde{\mu}_1^{(n)} - \tilde{\mu}_1)^q}{q!} \right) \mathcal{D}\boldsymbol{\Sigma} d\tilde{z} \right| \tag{A.117}$$

$$= 2 K_1^b K_2^{k-b-1} K_{\phi'} \mathcal{A}(k, I_i, I_\mu, b) \left(1 + \mathcal{O}\left(\frac{1}{n} \right) \right)$$
$$\cdot \left| \int \frac{\sigma_w \sigma_x}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \int \left(\sum_{q=1} \phi^{(q+1)}(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1} \tilde{z}) \frac{(\sigma_w \sigma_x \tilde{z})}{2\sqrt{\tilde{\mu}_1}} \frac{(\tilde{\mu}_1^{(n)} - \tilde{\mu}_1)^q}{q!} \right) \mathcal{D}\boldsymbol{\Sigma} d\tilde{z} \right| \tag{A.118}$$

$$= 2 K_1^b K_2^{k-b-1} K_{\phi'} \mathcal{A}(k, I_i, I_\mu, b) \left(1 + \mathcal{O}\left(\frac{1}{n} \right) \right)$$
$$\cdot \left| \int \frac{\sigma_w \sigma_x}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \left(\sum_{q=1} \phi^{(q+1)}(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1} \tilde{z}) \frac{(\sigma_w \sigma_x \tilde{z})}{2\sqrt{\tilde{\mu}_1}} \frac{\int (\tilde{\mu}_1^{(n)} - \tilde{\mu}_1)^q \mathcal{D}\boldsymbol{\Sigma}}{q!} \right) d\tilde{z} \right| \tag{A.119}$$

$$= 2K_1^b K_2^{k-b-1} K_{\phi'} \mathcal{A}(k, I_i, I_\mu, b) \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right)$$

$$\cdot \left| \int \frac{\sigma_w \sigma_x}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \left(\sum_{\substack{q=2\xi \\ \xi>1}} \phi^{(q+1)}(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1}\tilde{z}) \frac{(\sigma_w \sigma_x \tilde{z})}{2\sqrt{\tilde{\mu}_1}} \frac{(q-1)!! \sigma_{\mathbf{\Sigma}}^q}{n^{q/2} q!} \right) d\tilde{z} \right| \tag{A.120}$$

$$\leq 2K_1^b K_2^{k-b-1} K_{\phi'} \mathcal{A}(k, I_i, I_\mu, b) \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \sum_{\substack{q=2\xi \\ \xi>1}} \frac{\sigma_{\mathbf{\Sigma}}^q}{n^{q/2} q!!}$$

$$\cdot \left| \int \frac{\sigma_w \sigma_x}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \left(\phi^{(q+1)}(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1}\tilde{z}) \frac{(\sigma_w \sigma_x \tilde{z})}{2\sqrt{\tilde{\mu}_1}} \right) d\tilde{z} \right| \tag{A.121}$$

$$= 2K_1^b K_2^{k-b-1} K_{\phi'} \mathcal{A}(k, I_i, I_\mu, b) \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \sum_{\substack{q=2\xi \\ \xi>1}} \frac{\sigma_{\mathbf{\Sigma}}^q}{n^{q/2} q!!}$$

$$\cdot \left| \int \frac{\sigma_w^3 \sigma_x^3}{2\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \phi^{(q+2)}(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1}\tilde{z}) d\tilde{z} \right| \tag{A.122}$$

$$= 2K_1^b K_2^{k-b-1} K_{\phi'} \mathcal{A}(k, I_i, I_\mu, b) \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \sum_{\substack{q=2\xi \\ \xi>1}} \frac{\sigma_{\mathbf{\Sigma}}^q}{n^{q/2} q!!} \frac{\sigma_w^2 \sigma_x^2}{2} K_{\phi^{(q+2)}} \tag{A.123}$$

Finally for the third absolute difference, the following holds

$$\left| \int \tilde{\theta}_2^{k-b} \mathcal{A}(k, I_i, I_\mu, b) \left(\tilde{\theta}_1^{(n)^b} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) - \tilde{\theta}_1^b \right) \mathcal{D}\mathbf{\Sigma} \right|$$

$$= \tilde{\theta}_2^{k-b} \mathcal{A}(k, I_i, I_\mu, b) \left| \int \left(\tilde{\theta}_1^{(n)^b} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) - \tilde{\theta}_1^b \right) \mathcal{D}\mathbf{\Sigma} \right| \tag{A.124}$$

$$\leq 2\tilde{\theta}_2^{k-b} \mathcal{A}(k, I_i, I_\mu, b) K_1^{b-1} \left| \int \left(\tilde{\theta}_1^{(n)} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) - \tilde{\theta}_1 \right) \mathcal{D}\mathbf{\Sigma} \right| \tag{A.125}$$

$$= 2\tilde{\theta}_2^{k-b} \mathcal{A}(k, I_i, I_\mu, b) K_1^{b-1} \left| \int \frac{\sigma_w \sigma_x}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \left(\phi^2(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1^{(n)}}\tilde{z}) - \phi^2(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1}\tilde{z}) \right) d\tilde{z} \mathcal{D}\mathbf{\Sigma} \right|$$
$$\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.126}$$

$$= 2\tilde{\theta}_2^{k-b} \mathcal{A}(k, I_i, I_\mu, b) K_1^{b-1} \left| \int \frac{\sigma_w \sigma_x}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \left(\phi^2(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1^{(n)}}\tilde{z}) - \phi^2(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1}\tilde{z}) \right.\right.$$

$$\cdot \left. \left. + \sum_{q=1} \frac{\partial^q}{\partial \tilde{\mu}_1^{(n)^q}} \left[\phi^2(\sigma_w \sigma_x \sqrt{\tilde{\mu}_1^{(n)}}\tilde{z}) \right|_{\tilde{\mu}_1^{(n)} = \tilde{\mu}_1} \frac{(\tilde{\mu}_1^{(n)} - \tilde{\mu}_1)^q}{q!} \right) d\tilde{z} \mathcal{D}\mathbf{\Sigma} \right| \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.127}$$

$$= 2\tilde{\theta}_2^{k-b}\mathcal{A}(k, I_i, I_\mu, b)K_1^{b-1}\left|\int \frac{\sigma_w\sigma_x}{\sqrt{2\pi}}e^{-\frac{\tilde{z}^2}{2}} \left(\sum_{\substack{q=2k \\ s.t.k>1}} \frac{\partial^q\left(\phi^2(\sigma_w\sigma_x\sqrt{\tilde{\mu}_1^{(n)}}\tilde{z})\right)}{\partial\tilde{\mu}_1^{(n)q}}\bigg|_{\tilde{\mu}_1^{(n)}=\tilde{\mu}_1} \frac{\sigma_{\boldsymbol{\Sigma}}^q}{n^{q/2}q!!}\right) d\tilde{z}\right|\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.128}$$

$$= 2\tilde{\theta}_2^{k-b}\mathcal{A}(k, I_i, I_\mu, b)K_1^{b-1}\sum_{\substack{q=2k \\ s.t.k>1}} \frac{\sigma_{\boldsymbol{\Sigma}}^q}{n^{q/2}q!!} \cdot\left|\int \frac{\sigma_w\sigma_x}{\sqrt{2\pi}}e^{-\frac{\tilde{z}^2}{2}} \left(\frac{\partial^q\left(\phi^2(\sigma_w\sigma_x\sqrt{\tilde{\mu}_1^{(n)}}\tilde{z})\right)}{\partial\tilde{\mu}_1^{(n)q}}\bigg|_{\tilde{\mu}_1^{(n)}=\tilde{\mu}_1}\right) d\tilde{z}\right|\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.129}$$

The results in (A.107)-(A.109), (A.112)-(A.123), and (A.124)-(A.129) show that the contribution of the three absolute differences in (A.106) is $\mathcal{O}(1/\sqrt{n})$. Which then proves that

$$|\mathfrak{m}_k^{(n)} - \mathfrak{m}_k| \leq \mathcal{O}\left(\frac{1}{\sqrt{n}}\right). \tag{A.130}$$

□

A.1.5 Side Theorems

Convergence to the empirical spectral measure The identified expression for the Stieltjes transform of the eigen-distribution of the matrix $\mathbf{M}$ in Corallary 3.2.5.1 cannot be used for its evaluation. However, thanks to its explicit dependence on the dimensionality of the problem, the Stieltjes expression above can be used to understand if there is a unique spectral-measure defining the moments. As it was shown in [6], this can be done by satisfying Carleman's condition [87]. Here we are going to follow the same structure.

Theorem A.1.5 ([87, Theorem 4.3]). *If a sequence of moments* $(\mathfrak{m}_k)_{k\in\mathbb{N}^+}$ *is positive semi-definite and*

$$\sum_{k=1}^{\infty} \mathfrak{m}_k^{-\frac{1}{2k}} = +\infty \tag{A.131}$$

then there is one unique representing measure supported on $[0, +\infty)$.

For a probability measure, ρ, the moments sequence is always positive semi-definite. The Hankel matrix $\mathbf{H}$ defined on the moment sequence

$$\mathbf{H} = \int_0^{\infty} \begin{bmatrix} \lambda \\ \lambda^2 \\ \vdots \\ \lambda^n \end{bmatrix} \begin{bmatrix} \lambda & \lambda^2 & \dots & \lambda^n \end{bmatrix} \rho(\lambda) d\lambda \tag{A.132}$$

is positive semi-definite as for any vector $\mathbf{v}$

$$\mathbf{v}^\top \mathbf{H} \mathbf{v} = \int_0^{\infty} \mathbf{v}^\top \begin{bmatrix} \lambda \\ \lambda^2 \\ \vdots \\ \lambda^n \end{bmatrix} \begin{bmatrix} \lambda & \lambda^2 & \dots & \lambda^n \end{bmatrix} \mathbf{v} \rho(\lambda) d\lambda = \int_0^{\infty} \left(\begin{bmatrix} \lambda & \lambda^2 & \dots & \lambda^n \end{bmatrix} \mathbf{v} \right)^2 \rho(\lambda) d\lambda \geq 0 \tag{A.133}$$

and this implies that the moment sequence is positive semi-definite [87, Theorem 3.12].

Further, the second condition is satisfied if $\mathfrak{m}_k < C^k$, since

$$\sum_{k=1}^{\infty} \mathfrak{m}_k^{-\frac{1}{2k}} > \sum_{k=1}^{\infty} C^{-\frac{1}{2}} = +\infty, \tag{A.134}$$

and this holds in our setting.

Lemma A.1.6. *For* $\boldsymbol{M} = \frac{1}{m} \boldsymbol{Y}\boldsymbol{Y}^\top$ *and* $\boldsymbol{Y} = \phi\left(\boldsymbol{W}\tilde{\boldsymbol{X}}\right)$ *as defined in Theorem 3.2.2, there is a constant* C *such that the moments satisfy*

$$\mathfrak{m}_k = \varphi^{1-k} \sum_{I_i, I_\mu = 1}^{k} \sum_{b=0}^{I_i + I_\mu + 1} \frac{\mathcal{A}(k, I_i, I_\mu, b) \tilde{\theta}_1^b \tilde{\theta}_2^{k-b} \varphi^{I_i} \psi^{I_\mu}}{2\pi((k-I_i)(k-I_\mu))^{\frac{1}{2}}} \left(\frac{e}{k-I_i}\right)^{k-I_1} \left(\frac{e}{k-I_\mu}\right)^{k-I_\mu} \tag{A.135}$$

$$\leq C^k. \tag{A.136}$$

Proof. In [6], it is shown that the graphs that we consider as admissible in our work are equivalent to the cactus graphs considered in [27]. It was shown that the number of cactus graphs patterns, $\Theta(k)$, generated with k vertices is independent from the number of identification and simple cycles, and there exist constants $\tau > 0$ and $\nu > 1$ such that $\Theta(k) \sim \frac{3\tau}{4\sqrt{\pi}} \frac{\nu^{k-3/2}}{k^{5/2}}$. Therefore, there is a constant $\bar{C}$ such that the number of patterns $\mathcal{B}(\overrightarrow{k}, I_i, I_\mu, b)$ that we can generate with $k = \sum_i \overrightarrow{k}_i$ vertices, I_i and I_μ identifications, and b cycles of dimension two, is upper bounded by $\bar{C}\Theta(k)$.

Specifically, there are constants $\bar{C} > 0$, $\tau > 0$, and $\nu > 1$ for which

$$\mathcal{B}(\overrightarrow{k}, I_i, I_\mu, b) < \bar{C} \frac{3\tau}{4\sqrt{\pi}} \frac{\nu^{k-3/2}}{k^{5/2}} \tag{A.137}$$

and consequently

$$\mathfrak{m}_k = \varphi^{1-k} \sum_{I_i, I_\mu = 1}^{k} \sum_{b=0}^{I_i + I_\mu + 1} \frac{\mathcal{A}(k, I_i, I_\mu, b) \tilde{\theta}_1^b \tilde{\theta}_2^{k-b} \varphi^{I_1} \psi^{I_\mu}}{2\pi((k-I_i)(k-I_\mu))^{\frac{1}{2}}} \left(\frac{e}{k-I_i}\right)^{k-I_1} \left(\frac{e}{k-I_\mu}\right)^{k-I_\mu} \tag{A.138}$$

$$= \varphi^{1-k} \sum_{I_i, I_\mu = 1}^{k} \sum_{b=0}^{I_i + I_\mu + 1} \frac{\tilde{\theta}_1^b \tilde{\theta}_2^{k-b} \varphi^{I_1} \psi^{I_\mu}}{2\pi((k-I_i)(k-I_\mu))^{\frac{1}{2}}} \left(\frac{e}{k-I_i}\right)^{k-I_1} \left(\frac{e}{k-I_\mu}\right)^{k-I_\mu} \cdot \left(\sum_{\overrightarrow{k}} \mathcal{B}(\overrightarrow{k}, I_i, I_\mu, b) \left(\prod_{c=1}^{1+I_i+I_\mu-b} \tilde{\mu}_{\tilde{k}_c} \right) \right) \tag{A.139}$$

$$\leq \varphi^{1-k} \sum_{I_i, I_\mu = 1}^{k} \sum_{b=0}^{I_i + I_\mu + 1} \frac{\tilde{\theta}_1^b \tilde{\theta}_2^{k-b} \varphi^{I_1} \psi^{I_\mu}}{2\pi((k-I_i)(k-I_\mu))^{\frac{1}{2}}} \left(\frac{e}{k-I_i}\right)^{k-I_1} \left(\frac{e}{k-I_\mu}\right)^{k-I_\mu} \cdot \left(\sum_{\overrightarrow{k}} \mathcal{B}(\overrightarrow{k}, I_i, I_\mu, b) \left(\max_{k \in \{1,\ldots,k\}} \tilde{\mu}_k \right)^{1+I_i+I_\mu-b} \right) \tag{A.140}$$

$$= \varphi^{1-k} \sum_{I_i, I_\mu = 1}^{k} \sum_{b=0}^{I_i + I_\mu + 1} \frac{\tilde{\theta}_1^b \tilde{\theta}_2^{k-b} \varphi^{I_1} \psi^{I_\mu}}{2\pi((k-I_i)(k-I_\mu))^{\frac{1}{2}}} \left(\frac{e}{k-I_i}\right)^{k-I_1} \left(\frac{e}{k-I_\mu}\right)^{k-I_\mu} \cdot \left(\max_{k \in \{1,\ldots,k\}} \tilde{\mu}_k \right)^{1+I_i+I_\mu-b} \sum_{\overrightarrow{k}} \mathcal{B}(\overrightarrow{k}, I_i, I_\mu, b) \tag{A.141}$$

$$\overset{(A.137)}{\leq} \varphi^{1-k} \sum_{I_i, I_\mu = 1}^{k} \sum_{b=0}^{I_i + I_\mu + 1} \frac{\tilde{\theta}_1^b \tilde{\theta}_2^{k-b} \varphi^{I_1} \psi^{I_\mu}}{2\pi((k-I_i)(k-I_\mu))^{\frac{1}{2}}} \left(\frac{e}{k-I_i}\right)^{k-I_1} \left(\frac{e}{k-I_\mu}\right)^{k-I_\mu}$$

$$\cdot \left(\max_{k\in\{1,\dots,k\}} \tilde{\mu}_k \right)^{1+I_i+I_\mu-b} \bar{C} \frac{3\tau}{4\sqrt{\pi}} \frac{\nu^{k-3/2}}{k^{5/2}} \tag{A.142}$$

$$= \varphi^{1-k} \frac{3\bar{C}\tau}{4\sqrt{\pi}} \frac{\nu^{k-3/2}}{k^{5/2}} \sum_{I_i,I_\mu=1}^{k} \frac{\varphi^{I_1}\psi^{I_\mu}}{2\pi((k-I_i)(k-I_\mu))^{\frac{1}{2}}} \left(\frac{e}{k-I_i}\right)^{k-I_1} \left(\frac{e}{k-I_\mu}\right)^{k-I_\mu}$$

$$\cdot \left(\sum_{b=0}^{I_i+I_\mu+1} \tilde{\theta}_1^b \tilde{\theta}_2^{k-b} \left(\max_{k\in\{1,\dots,k\}} \tilde{\mu}_k \right)^{1+I_i+I_\mu-b} \right) \tag{A.143}$$

$$\leq \varphi^{1-k} \frac{3\bar{C}\tau}{4\sqrt{\pi}} \frac{\nu^{k-3/2}}{k^{5/2}} \sum_{I_i,I_\mu=1}^{k} \frac{\varphi^{I_1}\psi^{I_\mu}}{2\pi((k-I_i)(k-I_\mu))^{\frac{1}{2}}} \left(\frac{e}{k-I_i}\right)^{k-I_1} \left(\frac{e}{k-I_\mu}\right)^{k-I_\mu}$$

$$\cdot (I_i+I_\mu+1) \max(\tilde{\theta}_1,\tilde{\theta}_2)^k \max\left(\left(\max_{k\in\{1,\dots,k\}} \tilde{\mu}_k \right)^{1+I_i+I_\mu}, \left(\max_{k\in\{1,\dots,k\}} \tilde{\mu}_k \right)^{0} \right) \tag{A.144}$$

$$\leq \varphi^{1-k} \frac{3\bar{C}\tau}{4\sqrt{\pi}} \frac{\nu^{k-3/2}}{k^{5/2}} (2k+1) \max(\tilde{\theta}_1,\tilde{\theta}_2)^k k \max(\varphi,\varphi^k) \max(\phi,\phi^k) \frac{e^2}{2\pi} \cdot$$

$$\cdot \max\left(\left(\max_{k\in\{1,\dots,k\}} \tilde{\mu}_k \right)^{1+2k}, \left(\max_{k\in\{1,\dots,k\}} \tilde{\mu}_k \right)^{0} \right) \tag{A.145}$$

$$= \varphi \frac{3\bar{C}\tau\nu^{-3/2}}{4\sqrt{\pi}} \frac{k(2k+1)}{k^{5/2}} \max(\varphi,\varphi^k) \max(\phi,\phi^k) \frac{e^2}{2\pi} \cdot$$

$$\cdot \max\left(\left(\max_{k\in\{1,\dots,k\}} \tilde{\mu}_k \right)^{1+2k}, \left(\max_{k\in\{1,\dots,k\}} \tilde{\mu}_k \right)^{0} \right) \tilde{C}^k \tag{A.146}$$

$$\leq C^k \tag{A.147}$$

where in (A.145) we upper-bounded the function $((e/k)^k)/\sqrt{2\pi k}$ with its maximum on the integers, i.e. $e/\sqrt{2\pi}$, considering that for 0 the indefinite form is assigned the value of 1, as reported in a footnote when introducing the approximation of $\#_1$ and $\#_2$ at the beginning of Section A.1.4. □

Contribution of non-admissible graphs The computation of the moments relies on the set of admissible graphs, as defined in Definition 3.1.1. However, among the addends in (3.41) alternative graphs might appear, see Figure A.1, but they do not contribute significantly to the moments of the eigen-distribution.

The number of non-admissible graphs with $2k$ vertices and c identifications is of the order $\mathcal{O}(n_0^{2k-c})$ as the admissible graphs. However, as shown in [72][Supplementary

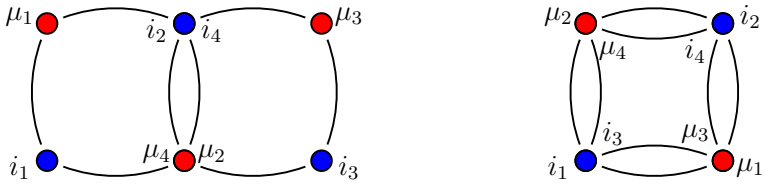


Figure A.1: Example of two non-admissible graphs with respectively $k = 6$ on the left, and $k = 4$ on the right.

Material 1.2.1] and [6][Section 3.1.4] the leading order contribution for non-admissible graphs is $\mathcal{O}(n_0^{c-1-k})$ rather than $\mathcal{O}(n_0^{c-k})$ as for the admissible graphs. This statement relies on the assumption that the activation function is such that $\left|\int \phi^k(\sigma_w\sigma_x z)\mathcal{D}z\right| < \infty$ since otherwise the contribution of a graph that consists in going through the same 2-dimensional cycle k-times could blow up. Therefore, non-admissible graphs generate a negligible contribution for $n \to \infty$.

A.2 Proofs Relevant to Theorem 3.3.1

A.2.1 Proof of Lemma 3.3.2

Proof. In the case where $k = 1$,

$$E_2^{(n)} = \int_{\mathbf{W},\mathbf{X}} \phi(\sum_l \mathbf{W}_{i_1,l}\mathbf{X}_{l,\mu_1})\phi(\sum_l \mathbf{W}_{i_1,l}\mathbf{X}_{l,\mu_1}) - \sigma_x^2 \sum_{p=1}^{n} \mathbf{W}_{i_1 p}\mathbf{W}_{i_1 p}\theta_3^2 \mathcal{D}\mathbf{W}\mathcal{D}\mathbf{X} \tag{A.148}$$

$$= \theta_1(1 + \mathcal{O}(1/n)) - \sigma_w^2\sigma_x^2\theta_3^2 \tag{A.149}$$

where θ_1 is found by following the same operations as in the proof of Lemma3.2.3 which lead to the following identity

$$\int_{\mathbf{W},\mathbf{X}} \phi(\sum_l \mathbf{W}_{i_1,l}\mathbf{X}_{l,\mu_1})\phi(\sum_l \mathbf{W}_{i_1,l}\mathbf{X}_{l,\mu_1})\mathcal{D}\mathbf{W}\mathcal{D}\mathbf{X}$$
$$= \int \frac{1}{\sqrt{2\pi}}\phi(\sigma_w\sigma_w z)^2 e^{-\frac{z^2}{2}} dz\,(1 + \mathcal{O}(1/n)) = \theta_1\,(1 + \mathcal{O}(1/n)) \tag{A.150}$$

and $\theta_3 = \int \phi'(\sigma_w\sigma_w z)e^{-\frac{z^2}{2}} dz$ □

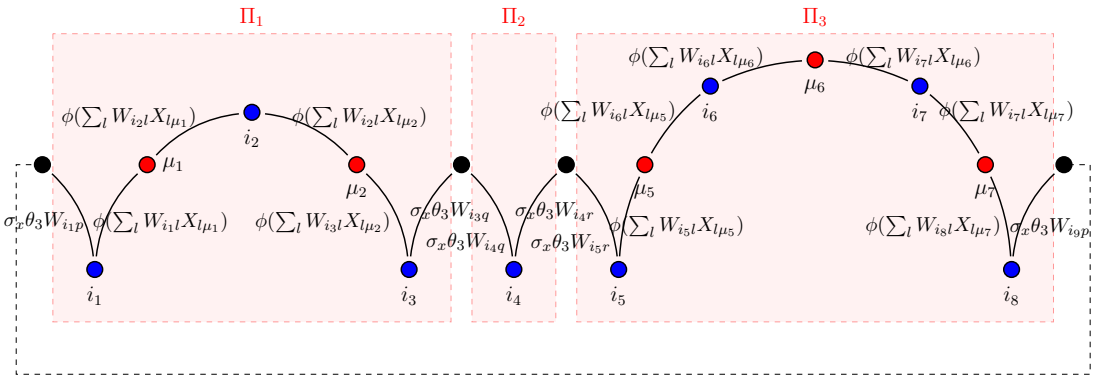


Figure A.2: Example of how in the Schur complement the contribution to the momentum is defined on as many independent blocks of products, Π_j, as the number of factors in the form $\sigma_x^2 \mathbf{W}_{i_\xi p} \mathbf{W}_{i_{\xi+1} p} \theta_3^2$.

A.2.2 Lemma 3.3.3

When considering $k > 1$ we need to consider all the mixed products in (3.124). As a matter of fact each term in the expansion of the product in E_{2k} consists of successions of products of the kind $\phi(\sum_l \mathbf{W}_{i_\xi, l} \mathbf{X}_{l,\mu_\xi}) \phi(\sum_l \mathbf{W}_{i_{\xi+1}, l} \mathbf{X}_{l,\mu_\xi})$ alternated with products of the kind $\sigma_x^2 \theta_3^2 \sum_{p=1}^n \mathbf{W}_{i_\xi p} \mathbf{W}_{i_{\xi+1} p}$. Each element of the kind $\sigma_x^2 \theta_3^2 \mathbf{W}_{i_\xi p} \mathbf{W}_{i_{\xi+1} p}$ divides the sequence of products into independent blocks Π_j. As we see in Figure A.2, with three factors of the kind $\sigma_x^2 \theta_3^2 \mathbf{W}_{i_\xi p} \mathbf{W}_{i_{\xi+1} p}$ we generate three independent blocks Π_j that are independent among themselves.

Lemma A.2.1. *For* $\boldsymbol{M} = \frac{1}{m} \mathbb{E}_{\boldsymbol{X}} \left[\boldsymbol{Y} \boldsymbol{Y}^\top \right] - \frac{1}{\sigma_x^2 m^2} \mathbb{E}_{\boldsymbol{x}} \left[\boldsymbol{Y} \boldsymbol{X}^\top \right] \mathbb{E}_{\boldsymbol{X}} \left[\boldsymbol{X} \ \boldsymbol{Y}^\top \right]$ *and* $\boldsymbol{Y} = \phi\left(\boldsymbol{W} \boldsymbol{X} \right)$ *as defined in Theorem 3.3.1, when* $k > 1$ *each independent block* Π_j^l*, with* l *terms* $\phi(\sum_l \boldsymbol{W}_{i_\xi, l} \boldsymbol{X}_{l,\mu_\xi}) \phi(\sum_l \boldsymbol{W}_{i_{\xi+1}, l} \boldsymbol{X}_{l,\mu_\xi})$*, generates the following contribution*

$$E^{(n)}_{\Pi_i^{(l)}} = -\frac{\sigma_w^2 \sigma_x^2 \theta_3^2 \theta_2^l}{n^{1+l}}. \tag{A.151}$$

with

$$\theta_2 = \left(\int \frac{\sigma_w \sigma_x}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} \phi'(\sigma_w \sigma_x \tilde{z}) d\tilde{z} \right)^2, \tag{A.152}$$

$$\theta_3 = \int_{z_1} \phi'\left(\sigma_x \sigma_w z_1 \right) \mathcal{D} z_1 \tag{A.153}$$

Proof. We focus on the integration of one independent block $\Pi_j^{(l)}$, where l identifies the numbers of factors $\phi(\sum_l \mathbf{W}_{i_\xi,l}\mathbf{X}_{l,\mu_\xi})\phi(\sum_l \mathbf{W}_{i_{\xi+1},l}\mathbf{X}_{l,\mu_\xi})$ between two of the $\sigma_x\theta_3\mathbf{W}_{i_\xi p}$ kind, i.e.

$$E^{(n)}_{\Pi_i^{(l)}} = -\int_{\mathbf{W},\mathbf{X}} \left\{ \sigma_x\theta_3\mathbf{W}_{i_1 p} \prod_{\xi=1}^{l} \left(\phi(\sum_l \mathbf{W}_{i_\xi,l}\mathbf{X}_{l,\mu_\xi})\phi(\sum_l \mathbf{W}_{i_{\xi+1},l}\mathbf{X}_{l,\mu_\xi}) \right) \sigma_x\theta_3\mathbf{W}_{i_{l+1}q} \right\} \mathcal{D}\mathbf{W}\mathcal{D}\mathbf{X}. \tag{A.154}$$

We are going to compute this expectation following the structure of the proof in [72]. We will introduce a dummy variable z with a delta Dirac function within each ϕ element, and then introduce a Fourier representation for all the arguments in the ϕ functions. We introduce for each factor Π_j the set $\mathcal{Z}_{\Pi_j} \subset \mathcal{Z}$ which contains only the combinations (i_ξ, μ_ν) that are included in the $\phi(\sum_l \mathbf{W}_{i_{\xi+1},l}\mathbf{X}_{l,\mu_\nu})$ arguments. Thus we consider auxiliary integrals over z, by adding delta functions enforcing $\mathbf{Z} = \mathbf{W}\mathbf{X}$ with

$$\mathbf{Z}_{i\mu} = \begin{cases} z_{i\mu} & \text{if } \ (i,\mu) \in \mathcal{Z}_{\Pi_j} \\ 0 & \text{otherwise.} \end{cases} \tag{A.155}$$

and consequently

$$E^{(n)}_{\Pi_i^{(l)}} = -\int \left\{ \sigma_x\theta_3\mathbf{W}_{i_1 p} \prod_{\xi=1}^{l} \left(\phi(\sum_l \mathbf{W}_{i_\xi,l}\mathbf{X}_{l,\mu_\xi})\phi(\sum_l \mathbf{W}_{i_{\xi+1},l}\mathbf{X}_{l,\mu_\xi}) \right) \sigma_x\theta_3\mathbf{W}_{i_{l+1}q} \right\} \mathcal{D}\mathbf{W}\mathcal{D}\mathbf{X} \tag{A.156}$$

$$= -\int \prod_{(\alpha,\beta)\in\mathcal{Z}_{\Pi_j}} \delta(z_{\alpha\beta} - \sum_k \mathbf{W}_{\alpha k}\mathbf{X}_{k\beta}) \left\{ \sigma_x^2\theta_3^2\mathbf{W}_{i_1 p} \prod_{\xi=1}^{l} \left(\phi(z_{i_\xi\mu_\xi})\phi(z_{i_\xi\mu_\xi}) \right) \mathbf{W}_{i_{l+1}q} \right\} \mathcal{D}\mathbf{W}\mathcal{D}\mathbf{X}\mathcal{D}z \tag{A.157}$$

$$= -\int e^{-itr\mathbf{\Lambda}^\top(\mathbf{W}\mathbf{X}-\mathbf{Z})} \sigma_x^2\theta_3^2\mathbf{W}_{i_1 p} \prod_{\xi=1}^{l} \left(\phi(z_{i_\xi\mu_\xi})\phi(z_{i_{\xi+1}\mu_\xi}) \right) \mathbf{W}_{i_{l+1}q} \mathcal{D}\mathbf{W}\mathcal{D}\mathbf{X}\mathcal{D}\lambda\mathcal{D}z \tag{A.158}$$

where

$$\mathcal{D}z = \prod_{(\alpha,\beta)\in\mathcal{Z}_{\Pi_j}} dz_{\alpha\beta} \tag{A.159}$$

and in the second equality we used the property that

$$\delta(x) = \frac{1}{2\pi}\int e^{i\lambda x} d\lambda \tag{A.160}$$

and therefore introduced the matrix $\mathbf{\Lambda} \in \mathbb{R}^{n \times m}$ whose entries are

$$\mathbf{\Lambda}_{i\mu} = \begin{cases} \lambda_{i\mu} & \text{if } (i,\mu) \in \mathcal{Z}_{\Pi_j} \\ 0 & \text{otherwise.} \end{cases} \tag{A.161}$$

and therefore

$$\mathcal{D}\lambda = \prod_{(\alpha,\beta)\in\mathcal{Z}_{\Pi_j}} \frac{d\lambda_{\alpha\beta}}{2\pi}. \tag{A.162}$$

We integrate over $\mathbf{X}$

$$\int e^{-itr(\mathbf{\Lambda}^\top \mathbf{W}\mathbf{X})}\mathcal{D}\mathbf{X} = \prod_{b,c=1}^{m,n} \int \frac{d\mathbf{X}_{cb}}{\sqrt{2\pi\sigma_x^2}} exp\left[-\frac{1}{2\sigma_x^2}\mathbf{X}_{cb}^2 - i\sum_{a=1}^{n}\lambda_{ab}\mathbf{W}_{ac}\mathbf{X}_{cb}\right] \tag{A.163}$$

$$= \prod_{b,c=1}^{m,n} \int \frac{d\mathbf{X}_{cb}}{\sqrt{2\pi\sigma_x^2}} exp\left[-\frac{1}{2\sigma_x^2}\left(\mathbf{X}_{cb} + i\sigma_x^2(\sum_{a=1}^{n}\lambda_{ab}\mathbf{W}_{ac})\right)^2 - \frac{\sigma_x^2}{2}(\sum_{a=1}^{n}\lambda_{ab}\mathbf{W}_{ac})^2\right] \tag{A.164}$$

$$= \prod_{b,c=1}^{m,n} exp\left[-\frac{\sigma_x^2}{2}(\sum_{a=1}^{n}\lambda_{ab}\mathbf{W}_{ac})^2\right] = exp\left[-\frac{\sigma_x^2}{2}\sum_{b,c=1}^{m,n}(\sum_{a=1}^{n}\lambda_{ab}\mathbf{W}_{ac})^2\right] \tag{A.165}$$

$$= exp\left[-\frac{\sigma_x^2}{2}\|\mathbf{\Lambda}^\top\mathbf{W}\|_F^2\right] = e^{-\frac{\sigma_x^2}{2}tr(\mathbf{W}^\top\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{W})}. \tag{A.166}$$

We also integrate over $\mathbf{W}$, first considering the case where $p \neq q$

$$\int_{\mathbf{W}} \sigma_x^2\theta_3^2 e^{-\frac{\sigma_x^2}{2}tr(\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{W}\mathbf{W}^\top)}\mathbf{W}_{i_1p}\mathbf{W}_{i_{l+1}q}\mathcal{D}\mathbf{W} \tag{A.167}$$

$$= \sigma_x^2\theta_3^2 \int \left(\prod_{i,j=1}^{n}\frac{d\mathbf{W}_{ij}}{\sqrt{2\pi\sigma_w^2/n}}\right) e^{-\frac{ntr(\mathbf{W}^\top\mathbf{W})}{2\sigma_w^2}} e^{-\frac{\sigma_x^2}{2}tr(\mathbf{W}^\top\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{W})}\mathbf{W}_{i_1p}\mathbf{W}_{i_{l+1}q} \tag{A.168}$$

$$= \sigma_x^2\theta_3^2 \int \left(\prod_{j=1}^{n}\frac{d^n\mathbf{w}_j}{(2\pi\sigma_w^2/n)^{n/2}}\right) e^{-\sum_{j=1}^{n}(\frac{n}{2\sigma_w^2}\mathbf{w}_j^\top\mathbf{w}_j + \frac{\sigma_x^2}{2}\mathbf{w}_j^\top\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{w}_j)}\mathbf{w}_p^{(i_1)}\mathbf{w}_q^{(i_{l+1})} \tag{A.169}$$

$$= \sigma_x^2\theta_3^2 \int \left(\prod_{\substack{j=1\\ j\neq p\\ j\neq q}}^{n}\frac{d^n\mathbf{w}_j}{(2\pi\sigma_w^2/n)^{n/2}}\right) e^{-\sum_{\substack{j=1\\ j\neq p\\ j\neq q}}^{n}(\frac{n}{2\sigma_w^2}\mathbf{w}_j^\top\mathbf{w}_j + \frac{\sigma_x^2}{2}\mathbf{w}_j^\top\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{w}_j)}$$
$$\cdot \underbrace{\int \frac{d^n\mathbf{w}_p}{(2\pi\sigma_w^2/n)^{n/2}}\mathbf{w}_p^{(i_1)} e^{-\frac{n}{2\sigma_w^2}\mathbf{w}_p^\top\mathbf{w}_p - \frac{\sigma_x^2}{2}\mathbf{w}_p^\top\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{w}_p}}_{=0}$$

$$\cdot \underbrace{\int \frac{d^n\mathbf{w}_p}{(2\pi\sigma_w^2/n)^{n/2}} \mathbf{w}_q^{(i_{l+1})} e^{-\frac{n}{2\sigma_w^2}\mathbf{w}_q^\top \mathbf{w}_q - \frac{\sigma_x^2}{2}\mathbf{w}_q^\top \mathbf{\Lambda}\mathbf{\Lambda}^\top \mathbf{w}_q}}_{=0} \tag{A.170}$$

$$= 0. \tag{A.171}$$

While now we will consider the case where $q = p$

$$\int_{\mathbf{W}} e^{-\frac{\sigma_x^2}{2} tr(\mathbf{\Lambda}\mathbf{\Lambda}^\top \mathbf{W}\mathbf{W}^\top)} \sigma_x^2\theta_3^2 \mathbf{W}_{i_1 p}\mathbf{W}_{i_{l+1}q}\mathcal{D}\mathbf{W}$$

$$= \sigma_x^2\theta_3^2 \int \left(\prod_{i,j=1}^{n} \frac{d\mathbf{W}_{ij}}{\sqrt{2\pi\sigma_w^2/n}}\right) e^{-\frac{ntr(\mathbf{W}^\top\mathbf{W})}{2\sigma_w^2}} e^{-\frac{\sigma_x^2}{2}tr(\mathbf{W}^\top\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{W})} \mathbf{W}_{i_1 p}\mathbf{W}_{i_{l+1}q} \tag{A.172}$$

$$= \sigma_x^2\theta_3^2 \int \left(\prod_{j=1}^{n} \frac{d^n\mathbf{w}_j}{(2\pi\sigma_w^2/n)^{n/2}}\right) e^{-\sum_{j=1}^{n}(\frac{n}{2\sigma_w^2}\mathbf{w}_j^\top\mathbf{w}_j + \frac{\sigma_x^2}{2}\mathbf{w}_j^\top\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{w}_j)} \mathbf{w}_p^{(i_1)}\mathbf{w}_q^{(i_{l+1})} \tag{A.173}$$

$$= \sigma_x^2\theta_3^2 \int \left(\prod_{\substack{j=1\\ j\neq p}}^{n} \frac{d^n\mathbf{w}_j}{(2\pi\sigma_w^2/n)^{n/2}}\right) e^{-\sum_{\substack{j=1\\ j\neq p}}^{n}(\frac{n}{2\sigma_w^2}\mathbf{w}_j^\top\mathbf{w}_j + \frac{\sigma_x^2}{2}\mathbf{w}_j^\top\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{w}_j)}$$
$$\cdot \int \frac{d^n\mathbf{w}_p}{(2\pi\sigma_w^2/n)^{n/2}} \mathbf{w}_p^{(i_1)}\mathbf{w}_p^{(i_{l+1})} e^{-\frac{n}{2\sigma_w^2}\mathbf{w}_p^\top\mathbf{w}_p - \frac{\sigma_x^2}{2}\mathbf{w}_p^\top\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{w}_p} \tag{A.174}$$

$$= \sigma_x^2\theta_3^2 \prod_{\substack{j=1\\ j\neq p}}^{n} \underbrace{\left(\int \frac{d^n\mathbf{w}_j}{(2\pi\sigma_w^2/n)^{n/2}} \frac{det^{-1}(\mathbf{I} + \frac{\sigma_x^2\sigma_w^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top)^{1/2}}{det^{-1}(\mathbf{I} + \frac{\sigma_x^2\sigma_w^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top)^{1/2}} e^{-\frac{n}{2\sigma_w^2}\mathbf{w}_j^\top\mathbf{w}_j - \frac{\sigma_x^2}{2}\mathbf{w}_j^\top\mathbf{\Lambda}\mathbf{\Lambda}^\top\mathbf{w}_j)}\right)}_{=\frac{1}{det(\mathbf{I}+\frac{\sigma_x^2\sigma_w^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top)^{1/2}}}$$
$$\cdot \int \frac{d^n\mathbf{w}_p}{(2\pi\sigma_w^2/n)^{n/2}} \frac{det^{-1}(\mathbf{I} + \frac{\sigma_x^2\sigma_w^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top)^{1/2}}{det^{-1}(\mathbf{I} + \frac{\sigma_x^2\sigma_w^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top)^{1/2}} \mathbf{w}_p^{(i_1)}\mathbf{w}_p^{(i_{l+1})} e^{-\frac{1}{2}\mathbf{w}_p^\top\left(\frac{n}{\sigma_w^2}\mathbf{I} + \sigma_x^2\mathbf{\Lambda}\mathbf{\Lambda}^\top\right)\mathbf{w}_p} \tag{A.175}$$

$$= \frac{\sigma_x^2\theta_3^2}{det(\mathbf{I} + \frac{\sigma_x^2\sigma_w^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top)^{n/2}} Cov\left[\mathbf{w}_p^{(i_1)}\mathbf{w}_p^{(i_{l+1})}\right]. \tag{A.176}$$

Consequently, each non-zero contribution to the momentum is going to be defined by addends in (3.124) whose elements in the form $\sigma_x\theta_3\mathbf{W}_{i_\xi p}$ have the same pedex p.

We notice that the covariance matrix of the vector w_j is

$$\left(\frac{n}{\sigma_w^2}\mathbf{I} + \sigma_x^2\mathbf{\Lambda}\mathbf{\Lambda}^\top\right)^{-1} = \frac{\sigma_w^2}{n}\sum_{k=0}^{\infty}(-\frac{\sigma_w^2\sigma_x^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top)^k \tag{A.177}$$

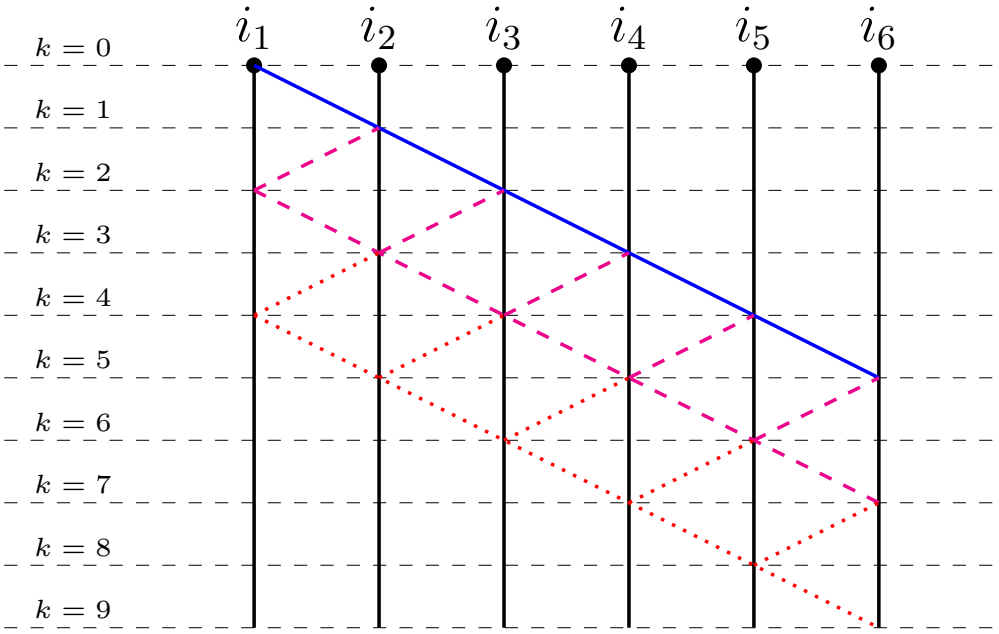


Figure A.3: Diagram of sequences of paths that allow to connect i_1 to i_{l+1}, with $l = 5$, for different number of steps k. The blue line corresponds to the unique path connecting i_0 to i_l with $k = l$ steps; the dashed magenta lines corresponds to the l paths connecting i_0 to i_l with $k = l + 2$ steps; and the dotted red lines corresponds to the l^2 paths connecting i_0 to i_l with $k = l + 4$ steps.

therefore assuming that n is large enough

$$Cov\left[\mathbf{w}_j^{(i_1)}\mathbf{w}_j^{(i_{l+1})}\right] = \left[\left(\frac{n}{\sigma_w^2}\mathbf{I} + \sigma_x^2\mathbf{\Lambda}\mathbf{\Lambda}^\top\right)^{-1}\right]_{i_1 i_{l+1}} = \left[\frac{\sigma_w^2}{n}\sum_{k=0}^{\infty}(-\frac{\sigma_w^2\sigma_x^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top)^k\right]_{i_1 i_{l+1}}. \tag{A.178}$$

Since $\mathbf{\Lambda}$ is defined on $\mathcal{Z}_{\Pi_j}$, the non-zero terms in $\left[(\mathbf{\Lambda}\mathbf{\Lambda}^\top)^k\right]_{i_1 i_l}$ correspond to products $\prod_{j=1}^{l}\left(\lambda_{i_j\mu_j}\lambda_{i_{j+1}\mu_j}\right)^{\tilde{l}_j}$. The sequence of exponents $\overrightarrow{l} = (\tilde{l}_1, \ldots, \tilde{l}_l)$ can be determined by considering diagrams as the one reported in Figure A.3 with $l = 5$. For each power k, the sequence of edges that are non-zero correspond to the paths that connect i_1 to i_l in exactly k steps; the exponent $\tilde{l}_j$ correspond to the number of times that the path goes through the interval (i_j, i_{j+1}).

Therefore, $k = l$ is the only case where all of the variables $\lambda_{i_j\mu_j}\lambda_{i_{j+1}\mu_j}$ have exponent of exactly one and this allows to write the covariance in the following way.

$$Cov\left[\mathbf{w}_j^{(i_1)}\mathbf{w}_j^{(i_{l+1})}\right] = \left[\frac{\sigma_w^2}{n}\sum_{k=0}^{\infty}(-\frac{\sigma_w^2\sigma_x^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top)^k\right]_{i_1 i_{l+1}} \tag{A.179}$$

$$= (-1)^l\frac{\sigma_w^2}{n}\prod_{j=1}^{l}\frac{\sigma_w^2\sigma_x^2}{n}\lambda_{i_j\mu_j}\lambda_{i_{j+1}\mu_j} + \sum_{\nu=1}^{\infty}(-1)^{l+\nu}\frac{\sigma_w^2}{n}\sum_{\substack{\vec{l}\ s.t.\\ \|\vec{l}\|_1=\nu+l\\ \tilde{l}_j\in\mathbb{N}_0}}\prod_{j=1}^{l}\left(\frac{\sigma_w^2\sigma_x^2}{n}\lambda_{i_j\mu_j}\lambda_{i_{j+1}\mu_j}\right)^{\tilde{l}_j} \tag{A.180}$$

Let's now consider the integral on $\mathcal{D}\lambda\mathcal{D}z$ of the first addend in (A.180) and denote it with $E^{(n)\ L.O.}_{\Pi_i^{(l)}}$. By considering $F(z) = \prod_{\xi=1}^{l}\left(\phi(z_{i_\xi\mu_\xi})\phi(z_{i_{\xi+1}\mu_\xi})\right)$

$$E^{(n)\ L.O.}_{\Pi_i^{(l)}} =$$

$$-\int \sigma_x^2\theta_3^2(-1)^l\frac{\sigma_w^2}{n}\left(\prod_{j=1}^{l}\frac{\sigma_w^2\sigma_x^2}{n}\lambda_{i_j\mu_j}\lambda_{i_{j+1}\mu_j}\right)\frac{1}{det(\mathbf{I}+\frac{\sigma_x^2\sigma_w^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top)^{n/2}}e^{-itr\mathbf{\Lambda}^\top\mathbf{Z}}F(z)\mathcal{D}\lambda\mathcal{D}z \tag{A.181}$$

$$= -\int \sigma_x^2\theta_3^2(-1)^l\frac{\sigma_w^2}{n}\left(\prod_{j=1}^{l}\frac{\sigma_w^2\sigma_x^2}{n}\lambda_{i_j\mu_j}\lambda_{i_{j+1}\mu_j}\right)e^{-\frac{n}{2}\log det(\mathbf{I}+\frac{\sigma_x^2\sigma_w^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top)}e^{-itr\mathbf{\Lambda}^\top\mathbf{Z}}F(z)\mathcal{D}\lambda\mathcal{D}z \tag{A.182}$$

$$= -(-1)^l\frac{\sigma_w^2\sigma_x^2\theta_3^2}{n}\int\left(\prod_{(\alpha,\beta)\in\mathcal{Z}_{\Pi_j}}\frac{\sigma_w\sigma_x}{\sqrt{n}}\lambda_{\alpha\beta}\right) e^{-\frac{\sigma_x^2\sigma_w^2}{2}tr(\mathbf{\Lambda}\mathbf{\Lambda}^\top)-\frac{n}{2}\sum_{\xi\geq 2}\frac{(-1)^{\xi+1}}{\xi}tr\left(\frac{\sigma_x^2\sigma_w^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top\right)^\xi - itr\mathbf{\Lambda}^\top\mathbf{Z}}F(z)\mathcal{D}\lambda\mathcal{D}z \tag{A.183}$$

where in the last equation we used the Taylor expansion $logdet|\mathbf{I}+\mathbf{X}| = \sum_{\xi=1}\frac{(-1)^{\xi+1}}{\xi}tr(\mathbf{X}^\xi)$. However, differently from the study of the covariance matrix $\mathbf{\Lambda}_{h^{(\ell)}}$, thanks to the $\mathbf{W}$ integral, we already have a factor containing all of the variables $\lambda_{\alpha\beta}$ exactly once. Therefore, we can consider only the first order expansion, ie. $\xi = 1$. Further we are also going to consider the following change of variable,

$$\tilde{\lambda}_{\alpha\beta} = \sigma_w\sigma_x\lambda_{\alpha\beta} \tag{A.184}$$

therefore

$$E_{\Pi_i^{(l)}}^{(n)\ L.O.} \approx -(-1)^l \frac{\sigma_w^2\sigma_x^2\theta_3^2}{n} \int \left(\prod_{(\alpha,\beta)\in\mathcal{Z}_{\Pi_j}} \frac{\sigma_w\sigma_x}{\sqrt{n}}\lambda_{\alpha\beta} \right) e^{-\frac{\sigma_x^2\sigma_w^2}{2}tr(\mathbf{\Lambda}\mathbf{\Lambda}^\top)-itr\mathbf{\Lambda}^\top\mathbf{Z}} F(z)\mathcal{D}\lambda\mathcal{D}z \tag{A.185}$$

$$= -(-1)^l \frac{\sigma_w^2\sigma_x^2\theta_3^2}{n} \int \left(\prod_{(\alpha,\beta)\in\mathcal{Z}_{\Pi_j}} \frac{\tilde{\lambda}_{\alpha\beta}}{\sqrt{n}} \right) e^{-\sum_{(\alpha,\beta)\in\mathcal{Z}_{\Pi_j}} \left(\frac{\tilde{\lambda}_{\alpha\beta}^2}{2} + i\frac{\tilde{\lambda}_{\alpha\beta}z_{\alpha\beta}}{\sigma_x\sigma_w} \right)} F(z) \left(\prod_{(\alpha,\beta)\in\mathcal{Z}_{\Pi_j}} \frac{d\tilde{\lambda}_{\alpha\beta}}{2\pi\sigma_w\sigma_x} \right) \left(\prod_{(\alpha,\beta)\in\mathcal{Z}_{\Pi_j}} dz_{\alpha\beta} \right) \tag{A.186}$$

$$= -(-1)^l \frac{\sigma_w^2\sigma_x^2\theta_3^2}{n} \prod_{(\alpha\beta)\in\mathcal{Z}_{\Pi_j}} \left(\int \frac{\tilde{\lambda}_{\alpha\beta}}{\sqrt{n}} e^{-\frac{\tilde{\lambda}_{\alpha\beta}^2}{2} - i\frac{\tilde{\lambda}_{\alpha\beta}z_{\alpha\beta}}{\sigma_x\sigma_w}} \phi(z_{\alpha\beta}) \frac{d\tilde{\lambda}_{\alpha\beta}}{2\pi\sigma_w\sigma_x} dz_{\alpha\beta} \right) \tag{A.187}$$

$$= -(-1)^l \frac{\sigma_w^2\sigma_x^2\theta_3^2}{n} \left(\int \frac{\tilde{\lambda}_{\alpha\beta}}{\sqrt{n}} e^{-\frac{1}{2}\left(\tilde{\lambda}_{\alpha\beta} + \frac{iz_{\alpha\beta}}{\sigma_x\sigma_w}\right)^2 - \frac{z_{\alpha\beta}^2}{2\sigma_x^2\sigma_w^2}} \phi(z_{\alpha\beta}) \frac{d\tilde{\lambda}_{\alpha\beta}}{2\pi\sigma_w\sigma_x} dz_{\alpha\beta} \right)^{2l} \tag{A.188}$$

$$= -(-1)^l \frac{\sigma_w^2\sigma_x^2\theta_3^2}{n} \left(\int \frac{\phi(z)}{\sqrt{2\pi}\sqrt{n}\sigma_w\sigma_x} e^{-\frac{z^2}{2\sigma_x^2\sigma_w^2}} \left(\int \frac{\tilde{\lambda}}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\tilde{\lambda} + \frac{iz}{\sigma_x\sigma_w}\right)^2} d\tilde{\lambda} \right) dz \right)^{2l} \tag{A.189}$$

$$= -(-1)^l \frac{\sigma_w^2\sigma_x^2\theta_3^2}{n} \left(\int \frac{\phi(z)}{\sqrt{2\pi}\sqrt{n}\sigma_w\sigma_x} e^{-\frac{z^2}{2\sigma_x^2\sigma_w^2}} \left(-\frac{iz}{\sigma_x\sigma_w} \right) dz \right)^{2l} \tag{A.190}$$

$$= -(-1)^l \frac{\sigma_w^2\sigma_x^2\theta_3^2}{n} \left(\frac{-1}{n} \right)^l \left(\int_{\mathbf{Z}} \frac{\phi(z)}{\sqrt{2\pi}} e^{-\frac{z^2}{2\sigma_x^2\sigma_w^2}} \frac{z}{\sigma_x^2\sigma_w^2} dz \right)^{2l}. \tag{A.191}$$

Now we introduce the following change of variable

$$\tilde{z} = \frac{z}{\sigma_x\sigma_w} \tag{A.192}$$

and then

$$E_{\Pi_i^{(l)}}^{(n)\ L.O.} = -\frac{\sigma_w^2\sigma_x^2\theta_3^2}{n^{1+l}} \left(\int_{\mathbf{Z}} \frac{\tilde{z}\phi(\sigma_x\sigma_w\tilde{z})}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} d\tilde{z} \right)^{2l} \tag{A.193}$$

$$= -\frac{\sigma_w^2\sigma_x^2\theta_3^2}{n^{1+l}} \left(\sigma_w\sigma_x \int_{\mathbf{Z}} \frac{\phi'(\sigma_x\sigma_w\tilde{z})}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} d\tilde{z} \right)^{2l} \tag{A.194}$$

$$= -\frac{\sigma_w^2 \sigma_x^2 \theta_3^2 \theta_2^l}{n^{1+l}}. \tag{A.195}$$

The integral on $\mathcal{D}\lambda\mathcal{D}z$ for the remaining addends in (A.180) lead to the computation of higher moments of the gaussian variable λ, which are finite. However, $1/n$ is elevated to the same power as λ and therefore the contribution of the remaining addends in (A.180) is $E_{\Pi_i^{(l)}}^{(n)\ L.O.}\mathcal{O}(1/n)$.

It then follows that

$$E_{\Pi_i^{(l)}}^{(n)} = E_{\Pi_i^{(l)}}^{(n)\ L.O.} + E_{\Pi_i^{(l)}}^{(n)\ L.O.}\left(\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.196}$$

$$= -\frac{\sigma_w^2 \sigma_x^2 \theta_3^2 \theta_2^l}{n^{1+l}}\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right). \tag{A.197}$$

□

A.2.2.1 Proof of Lemma 3.3.3

Proof. Now we can use Lemma A.2.1 for the computation of E_{2k}. Let's assume $k > 1$, and compute the contribution of each term ω in the expansion of (3.124). Specifically each term has an associated number n_w of independent blocks Π_j, and we denote with $n_\phi^{(\xi)}$ and $n_w^{(\phi)}$ the number of factors of the type $\phi(\sum_l \mathbf{W}_{i_\xi,l}\mathbf{X}_{l,\mu_\xi})\phi(\sum_l \mathbf{W}_{i_{\xi+1},l}\mathbf{X}_{l,\mu_\xi})$ and $\sigma_x^2\theta_3^2\mathbf{W}_{i_\xi p}\mathbf{W}_{i_{\xi+1}p}$ respectively within each block Π_ξ, note that $k = \sum_\xi (n_w^{(\xi)} + n_\phi^{(\xi)})$, and $n_w^{(\xi)} = 1$ for each block. Further, we also fix the index p in the factors $\sigma_x^2\theta_3^2\mathbf{W}_{i_\xi p}\mathbf{W}_{i_{\xi+1}p}$, and then

$$E_\omega^{(k,n_w,n_\phi,p)} = \int_{\mathbf{W},\mathbf{X}} \prod_{\xi=1}^{n_w}\left(\sigma_x\theta_3\mathbf{W}_{i_\xi p}\left(\prod_{\ell=1}^{n_\phi^{(\xi)}}\phi(\sum_l \mathbf{W}_{i_{\xi+\ell},l}\mathbf{X}_{l,\mu_{\xi+\ell}})\phi(\sum_l \mathbf{W}_{i_{\xi+\ell+1},l}\mathbf{X}_{l,\mu_{\xi+\ell}})\right)\right.$$
$$\left.\sigma_x\theta_3\mathbf{W}_{i_{\xi+1}p}\right)\mathcal{D}\mathbf{W}\mathcal{D}\mathbf{X} \tag{A.198}$$

$$= \prod_{\xi=1}^{n_w} E_{\Pi_\xi^{\left(n_\phi^{(\xi)}\right)}} = \prod_{\xi=1}^{n_w}\frac{-\sigma_w^2\sigma_x^2\theta_3^2\theta_2^{n_\phi^{(\xi)}}}{n^{1+n_\phi^{(\xi)}}}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.199}$$

$$= \left(\frac{-\sigma_w^2\sigma_x^2\theta_3^2}{n}\right)^{n_w}\left(\frac{\theta_2}{n}\right)^{\sum_{\xi=1}^{n_w} n_\phi^{(\xi)}}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.200}$$

$$= \left(-\frac{\sigma_w^2\sigma_x^2\theta_3^2}{n}\right)^{n_w}\left(\frac{\theta_2}{n}\right)^{k-n_w}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.201}$$

The expected contribution for the $2k$ cycle is defined by considering the contribution $\tilde{E}^{\phi}_{2k} = n_0^{1-k}\theta_2^k\big(1+\mathcal{O}\left(\frac{1}{n}\right)\big)$ of the addend ω whose $n_w = 0$ and of the addends for which $n_w \neq 0$

$$E^{(n)}_{2k} = \tilde{E}^{\phi}_{2k} + \sum_{n_w=1}^{k}\binom{k}{n_w}\sum_{p=1}^{n} E_{\omega}^{(k,n_w,n_\phi,p)} \tag{A.202}$$

$$= n_0^{1-k}\theta_2^k\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) + n_0^{1-k}\sum_{n_w=1}^{k}\binom{k}{n_w}\left(-\sigma_w^2\sigma_x^2\theta_3^2\right)^{n_w}\theta_2^{k-n_w}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.203}$$

$$= n_0^{1-k}\left(\theta_2 - \sigma_w^2\sigma_x^2\theta_3^2\right)^k\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.204}$$

□

A.2.3 Proof of Proposition 3.3.4

For each possible pattern we have to consider how many graphs, $\#_1$ and $\#_2$ we can build by varying respectively the I_i i-identifications, and the I_μ μ-identifications. Since, we have to assign for each of the $k - I_i$ i-identification a value among the n_1 available without reinsertion

$$\#_1 = \binom{n_1}{k-I_i} = \left(\frac{n_1 e}{k-I_i}\right)^{k-I_1}(2\pi(k-I_i))^{-1/2}(1+o(\frac{1}{n_1})) \tag{A.205}$$

and equivalently

$$\#_2 = \binom{m}{k-I_\mu} = \left(\frac{me}{k-I_\mu}\right)^{k-I_\mu}(2\pi(k-I_\mu))^{-1/2}(1+o(\frac{1}{m})). \tag{A.206}$$

Thanks to the formulation for the contribution of $2k$-cycles in Lemmas 3.3.2 and 3.3.3, we are now able to determine the contribution of any admissible graph G.

Proposition A.2.2. *For* $\boldsymbol{M} = \frac{1}{m}\mathbb{E}_{\boldsymbol{X}}\left[\boldsymbol{Y}\boldsymbol{Y}^\top\right] - \frac{1}{\sigma_x^2 m^2}\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{Y}\boldsymbol{X}^\top\right]\mathbb{E}_{\boldsymbol{X}}\left[\boldsymbol{X}\ \boldsymbol{Y}^\top\right]$ *and* $\boldsymbol{Y} = \phi\left(\boldsymbol{W}\boldsymbol{X}\right)$ *as defined in Theorem 3.3.1, an admissible graph* G *with* $2k$ *edges,*

b blocks of size 1, c blocks of dimensions $\overrightarrow{k} = \{k_1, \dots, k_c\}$*, and* b^μ *of the* I_μ *identifications define a one dimensional cycle, while the remaining* c^μ *of the* I_μ *identifications define larger cycles, is such that the contribution of the graph,* $E_G^{(n)}$*, grows as*

$$E_G^{(n)} = n_0^{c-\sum_i k_i} \left(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2\right)^{\sum_i k_i} \left(\theta_1 - \sigma_x^2\sigma_w^2\theta_3^2\right)^{b^i+b^\mu} \left[1 + \frac{2\sigma_x^2\sigma_w^2\theta_3^2\theta_2 - \sigma_x^4\sigma_w^4\theta_3^4}{(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)^2}\right]^{c^\mu} \left[1 + \frac{\theta_1\sigma_x^2\sigma_w^2\theta_3^2 + \sigma_x^2\sigma_w^2\theta_3^2\theta_2 - \sigma_x^4\sigma_w^4\theta_3^4}{(\theta_1 - \sigma_x^2\sigma_w^2\theta_3^2)(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)}\right]^{b^\mu} \cdot \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.207}$$

Proof. When considering the contribution of a generic composite graph we have to keep into consideration what kind of identification defines it. Specifically, if an identification is of the i-type, then this allows to consider the two connected subgraphs separately. However, if it is a μ-type contribution, then there is a correction term that has to be kept into consideration.

To give an intuition on why the correction term arises, and to introduce the cases that we have to investigate, we are going to consider a graph G generated by two cycles. When they are connected by an i-identification, the factors $\sigma_x^2\theta_3^2\mathbf{W}_{i_\xi p}\mathbf{W}_{i_{\xi+1}q}$ always appear as a pair in each constituent cycle of the graph. However, if they are connected with a μ-identification, this is not the case; the factors $\sigma_x^2\theta_3^2\mathbf{W}_{i_\xi p}\mathbf{W}_{i_{\xi+1}q}$ can be split by the identification in the two different cycles. We will show that if none of the four edges connected to the identification vertex are of the $\sigma_x\theta_3\mathbf{W}_{i_\xi p}$ kind, then the cycles can be considered separately, while if not, they have to be considered as one. Considering separately two cycles implies that it is possible to choose the column p of the matrices $\mathbf{w}$ independently, thus leading to a n_0 fold increase in the estimation.

Now we are going to consider the expected contribution of a general graph G, which is defined by the following integral

$$E_G^{(n)} = \int \prod_{\xi=1}^{k} \left(\phi(\sum_l \mathbf{W}_{i_\xi,l}\mathbf{X}_{l,\mu_\xi})\phi(\sum_l \mathbf{W}_{i_{\xi+1},l}\mathbf{X}_{l,\mu_\xi}) - \sigma_x^2 \sum_{p=1}^{n} \mathbf{W}_{i_\xi p}\mathbf{W}_{i_{\xi+1}p}\theta_3^2\right) \mathcal{D}\mathbf{W}\mathcal{D}\mathbf{X}. \tag{A.208}$$

We are going to compute the contribution $E_G^{(n)}$ by considering each term, ω, in the resulting sum that results from the expansion of (A.208). However, to simplify the

computations, we are first going to consider the contribution of the different kinds of blocks of dependent variable Π_j that determine the addend ω. As in Lemma A.2.1, if there are no identifications between two successive terms $\sigma_x^2\theta_3^2\mathbf{W}_{i_\xi p}\mathbf{W}_{i_{\xi+1}p}$ then a block is isolated to the factors of the type $\phi(\sum_l \mathbf{W}_{i_\xi,l}\mathbf{X}_{l,\mu_\xi})\phi(\sum_l \mathbf{W}_{i_{\xi+1},l}\mathbf{X}_{l,\mu_\xi})$ between two of the type $\sigma_x\theta_3\mathbf{W}_{i_\xi p}$ and their contribution is the same as in Lemma A.2.1. We refer to the contribution of these terms as ${}_0E^{(n)}_{\Pi_j}$.

Differently from the cycle case, if there are some identifications the blocks might be defined by more than two $\sigma_x\theta_3\mathbf{W}_{i_\xi p}$ factors. We consider the following three kind of identifications:

- in the block Π_j there are only two factors of the type $\sigma_x\theta_3\mathbf{W}_{i_\xi p}$ and there are some identifications on the vertices in between the two blocks, see Figure A.4a.
- there are multiple factors $\sigma_x^2\theta_3^2\mathbf{W}_{i_\xi p}\mathbf{W}_{i_{\xi+1}q}$ defining the block Π_j, but none is such that μ_ξ is an identification, see Figure A.4b as an example;
- there is at least one factor $\sigma_x^2\theta_3^2\mathbf{W}_{i_\xi p}\mathbf{W}_{i_{\xi+1}q}$ such that μ_ξ is an identification, see Figure A.4c as an example. This corresponds to the term $\sigma_x^2\theta_3^2\mathbf{W}_{i_\xi p}\mathbf{W}_{i_{\xi+1}q}$ being split into two different cycles.

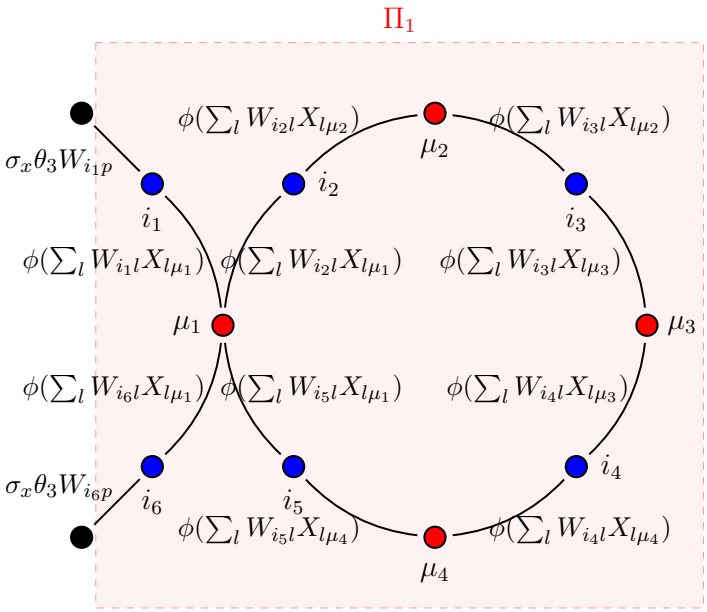
We are now going to identify all of their contributions.

First Kind: Let's consider the integration over the block $\Pi_j^{(l,p)}$ with l factors $\phi(\sum_l \mathbf{W}_{i_\xi,l}\mathbf{X}_{l,\mu_\xi})\phi(\sum_l \mathbf{W}_{i_{\xi+1},l}\mathbf{X}_{l,\mu_\xi})$ and p specifying the column considered of $\mathbf{W}$[2]. Then

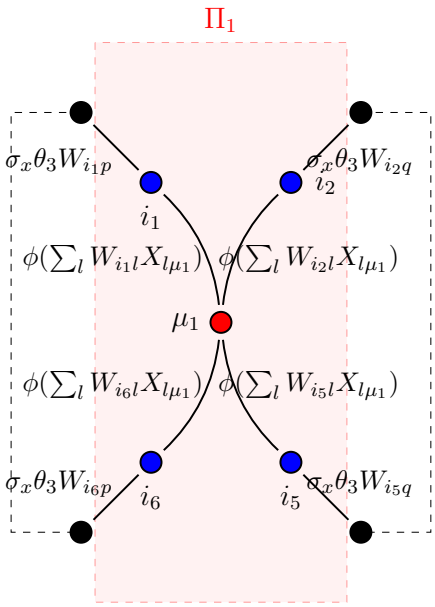
$$ {}_IE^{(n)}_{\Pi_i^{(l,p)}} = -\int \sigma_x\theta_3\mathbf{W}_{i_1 p}\prod_{\xi=1}^{l}\left(\phi(\sum_l \mathbf{W}_{i_\xi,l}\mathbf{X}_{l,\mu_\xi})\phi(\sum_l \mathbf{W}_{i_{\xi+1},l}\mathbf{X}_{l,\mu_\xi})\right)\sigma_x\theta_3\mathbf{W}_{i_{l+1}p}\mathcal{D}\mathbf{W}\mathcal{D}\mathbf{X}. \tag{A.209}$$

and we define n_{ind} as the number of identifications within the block; this corresponds to the number of complete cycles between the two factors $\sigma_x\theta_3\mathbf{W}_{i_1 p}$ and $\sigma_x\theta_3\mathbf{W}_{i_{l+1}p}$. By integrating over $\mathbf{X}$ and $\mathbf{W}$ as done in Lemma A.2.1 to retrieve

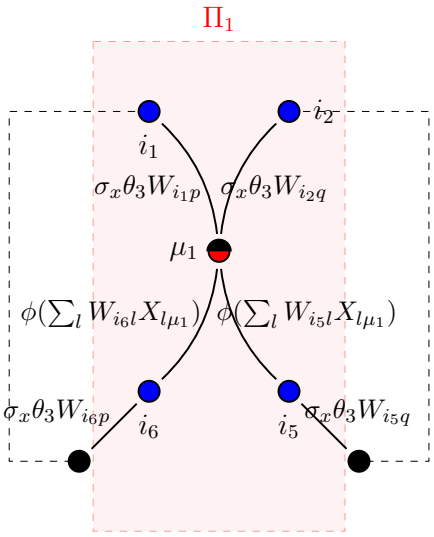
[2]If the two factors were relative to different columns p and q the contribution would be null



(a) First kind of identification



(b) Second kind of identification



(c) Third kind of identification

Figure A.4: Illustration of some examples of how the $\sigma_x^2\theta_3^2\mathbf{W}_{i_\xi p}\mathbf{W}_{i_{\xi+1}q}$ factors might be arranged around some identifications.

(A.176), we confirm that the two factors $\mathbf{W}_{i_1p}$ and $\mathbf{W}_{i_{l+1}p}$ must be relative to the same column and thus

$$_IE^{(n)}_{\Pi_i^{(l,p)}} = -\int \sigma_x^2\theta_3^2(-1)^l\frac{\sigma_w^2}{n}Cov\left[\mathbf{w}_j^{i_0}\mathbf{w}_j^{(i_{l+1})}\right]\frac{1}{det(\mathbf{I}+\frac{\sigma_x^2\sigma_w^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top)^{n/2}}e^{-itr\mathbf{\Lambda}^\top\mathbf{Z}}F(z)\mathcal{D}\lambda\mathcal{D}z \tag{A.210}$$

Let's now denote the set of indices that are respectively on the shortest path between the two factors $\sigma_x\theta_3\mathbf{W}_{i_1p}$ and $\sigma_x\theta_3\mathbf{W}_{i_{l+1}p}$ or in the ith cycle in Π_j as $\mathfrak{D}$ and $\mathfrak{C}_i$. Then, by referring to all of the possible combinations of the different $\mathfrak{C}_i$ as $\pi(\mathfrak{C}_i)$, the covariance is

$$Cov\left[\mathbf{w}_j^{i_0}\mathbf{w}_j^{(i_{l+1})}\right] = \left[\left(\frac{n}{\sigma_w^2}\mathbf{I}+\sigma_x^2\mathbf{\Lambda}\mathbf{\Lambda}^\top\right)^{-1}\right]_{i_0i_{l+1}} = \left[\frac{\sigma_w^2}{n}\sum_{k=0}^{\infty}(-\frac{\sigma_w^2\sigma_x^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top)^k\right]_{i_0i_{l+1}} \tag{A.211}$$

$$= \frac{\sigma_w^2}{n}\sum_{k=l}^{\infty}\left[(-\frac{\sigma_w^2\sigma_x^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top)^k\right]_{i_0i_{l+1}} \tag{A.212}$$

$$\begin{aligned}&= (-1)^{|\mathfrak{D}|}\frac{\sigma_w^2}{n}\prod_{(\alpha\beta)\in\mathfrak{D}}\frac{\sigma_w\sigma_x}{\sqrt{n}}\lambda_{\alpha\beta}+\sum_{\mathfrak{C}\in\pi(\mathfrak{C}_i)}(-1)^{|\mathfrak{D}|+|\mathfrak{C}|}\frac{\sigma_w^2}{n}\prod_{(\alpha\beta)\in\mathfrak{D}\cup\mathfrak{C}}\frac{\sigma_w\sigma_x}{\sqrt{n}}\lambda_{\alpha\beta}\\ &+\sum_{\nu=1}^{\infty}(-1)^{l+\nu}\frac{\sigma_w^2}{n}\left(\prod_{(\gamma,\delta)\in\mathfrak{D}}\left(\frac{\sigma_w\sigma_x}{\sqrt{n}}\lambda_{\gamma\delta}\right)^{\tilde{l}_j}\right)\sum_{\substack{\vec{l}\ s.t.\\ \|\vec{l}\|_1=\nu\\ \tilde{l}_j\in\mathbb{N}_0}}\prod_{(\alpha,\beta)\in\cup_i\mathfrak{C}_i\cup\mathfrak{D}}\left(\frac{\sigma_w\sigma_x}{\sqrt{n}}\lambda_{\alpha\beta}\right)^{\tilde{l}_j}.\end{aligned} \tag{A.213}$$

where $|\mathfrak{C}|$ and $|\mathfrak{D}|$ correspond to the cardinality of the two set, hence twice the number of factors $\phi(\sum_l \mathbf{W}_{i_\xi,l}\mathbf{X}_{l,\mu_\xi})\phi(\sum_l \mathbf{W}_{i_{\xi+1},l}\mathbf{X}_{l,\mu_\xi})$ identified in the sets.

Because of the same argument as in Lemma A.2.1, the terms of the covariance where the variables $\lambda_{\alpha\beta}$ have exponential at maximum equal to one are the leading contribution of $E^{(n)}_{\Pi_j^{(l,p)}}$ up to a term $\mathcal{O}(1/n)$. Among these terms we now show that the leading contribution in the covariance is given by the term just relying on $\mathfrak{D}$; each time a cycle is considered, the contribution incurs a cost of $1/n$. By approximating the covariance with the contribution of the shortest path between

the two factors of the $\sigma_x\theta_3\mathbf{W}_{i_{l+1}p}$ kind, we find the following contribution

$$
\begin{aligned}
&-\int \sigma_x^2\theta_3^2(-1)^{|\mathfrak{D}|}\frac{\sigma_w^2}{n}\left(\prod_{(\alpha\beta)\in\mathfrak{D}}^{|\mathfrak{D}|}\frac{\sigma_w\sigma_x}{\sqrt{n}}\lambda_{\alpha\beta}\right)\frac{1}{det(\mathbf{I}+\frac{\sigma_x^2\sigma_w^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top)^{n/2}}e^{-itr\mathbf{\Lambda}^\top\mathbf{Z}}F(z)\mathcal{D}\lambda\mathcal{D}z\\
&=-\int \sigma_x^2\theta_3^2(-1)^{|\mathfrak{D}|}\frac{\sigma_w^2}{n}\left(\prod_{(\alpha\beta)\in\mathfrak{D}}^{|\mathfrak{D}|}\frac{\sigma_w\sigma_x}{\sqrt{n}}\lambda_{\alpha\beta}\right)e^{-\frac{n}{2}\log det(\mathbf{I}+\frac{\sigma_x^2\sigma_w^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top)}e^{-itr\mathbf{\Lambda}^\top\mathbf{Z}}F(z)\mathcal{D}\lambda\mathcal{D}z
\end{aligned}
\tag{A.214}
$$

$$
\begin{aligned}
&=-(-1)^{|\mathfrak{D}|}\frac{\sigma_w^2\sigma_x^2\theta_3^2}{n}\int\left(\prod_{(\alpha\beta)\in\mathfrak{D}}^{|\mathfrak{D}|}\frac{\sigma_w\sigma_x}{\sqrt{n}}\lambda_{\alpha\beta}\right)\\
&\qquad e^{-\frac{\sigma_x^2\sigma_w^2}{2}tr(\mathbf{\Lambda}\mathbf{\Lambda}^\top)-\frac{n}{2}\sum_{\xi\geq 2}\frac{(-1)^{\xi+1}}{\xi}tr\left(\frac{\sigma_x^2\sigma_w^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top\right)^\xi-itr\mathbf{\Lambda}^\top\mathbf{Z}}F(z)\mathcal{D}\lambda\mathcal{D}z
\end{aligned}
\tag{A.215}
$$

In the same way as in Preposition A.1.4, the leading order of the contribution ${}_IE^{(n)}_{\Pi_i^{(l,\mathbf{p})}}$ is defined with a relative error $\mathcal{O}(\frac{1}{n})$ by considering the different cycles separately. Moreover, as for Lemma A.2.1 there is a relative error of $\mathcal{O}(1/n)$ by considering only the zeroth order approximation of the exponent $e^{\frac{n}{2}\sum_{\xi\geq 2}\frac{(-1)^{\xi+1}}{\xi}tr\left(\frac{\sigma_x^2\sigma_w^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top\right)^\xi}$ when considering the direct arch $\mathcal{D}$ and first order for the remaining cycles. Therefore

$$
\begin{aligned}
&-(-1)^{|\mathfrak{D}|}\frac{\sigma_w^2\sigma_x^2\theta_3^2}{n}\int\left(\prod_{(\alpha\beta)\in\mathfrak{D}}^{|\mathfrak{D}|}\frac{\sigma_w\sigma_x}{\sqrt{n}}\lambda_{\alpha\beta}\right)F(z)\\
&\quad exp\left\{-\sum_{\alpha\beta\in\mathfrak{D}\cup_{i=1}^{n_{ind}}\mathfrak{C}_i}\left(\frac{\sigma_x^2\sigma_w^2}{2}\lambda_{\alpha\beta}^2-i\lambda_{\alpha\beta}z_{\alpha\beta}\right)-\frac{n}{2}\sum_{\xi\geq 2}\frac{(-1)^{\xi+1}}{\xi}tr\left(\frac{\sigma_x^2\sigma_w^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top\right)^\xi\right\}\mathcal{D}\lambda\mathcal{D}z\\
&=-(-1)^{|\mathfrak{D}|}\frac{\sigma_w^2\sigma_x^2\theta_3^2}{n}\int\left(\prod_{(\alpha\beta)\in\mathfrak{D}}^{|\mathfrak{D}|}\frac{\sigma_w\sigma_x}{\sqrt{n}}\lambda_{\alpha\beta}\right)F(z)\\
&\quad exp\left\{-\sum_{\alpha\beta\in\mathfrak{D}\cup_{i=1}^{n_{ind}}\mathfrak{C}_i}\left(\frac{\sigma_x^2\sigma_w^2}{2}\lambda_{\alpha\beta}^2-i\lambda_{\alpha\beta}z_{\alpha\beta}\right)-\frac{n}{2}\sum_{i=1}^{n_{ind}}\frac{(-1)^{|\mathfrak{C}_i|+1}}{|\mathfrak{C}_i|}\prod_{(\alpha\beta)\in\mathfrak{C}_i}\frac{\sigma_x\sigma_w}{\sqrt{n}}\lambda_{\alpha\beta}\right\}\mathcal{D}\lambda\mathcal{D}z\\
&\qquad\cdot\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)
\end{aligned}
\tag{A.216}
$$

$$= -(-1)^{|\mathfrak{D}|}\frac{\sigma_w^2\sigma_x^2\theta_3^2}{n}\int\left(\prod_{(\alpha\beta)\in\mathfrak{D}}^{|\mathfrak{D}|}\frac{\sigma_w\sigma_x}{\sqrt{n}}\lambda_{\alpha\beta}\phi(z_{\alpha\beta})e^{-\sum_{\alpha\beta\in\mathfrak{D}}\left(\frac{\sigma_x^2\sigma_w^2}{2}\lambda_{\alpha\beta}^2-i\lambda_{\alpha\beta}z_{\alpha\beta}\right)}\right)$$
$$\left(\prod_{i=1}^{n_{ind}}(\prod_{\alpha\beta\in\mathfrak{C}_i}\phi(z_{\alpha\beta}))e^{-\sum_{\alpha\beta\in\mathfrak{C}_i}\left(\frac{\sigma_x^2\sigma_w^2}{2}\lambda_{\alpha\beta}^2-i\lambda_{\alpha\beta}z_{\alpha\beta}\right)+\frac{n}{2}\frac{(-1)^{|\mathfrak{C}_i|}}{|\mathfrak{C}_i|}2|\mathfrak{C}_i|\prod_{(\alpha\beta)\in\mathfrak{C}_i}\frac{\sigma_x\sigma_w}{\sqrt{n}}\lambda_{\alpha\beta}}\right)\mathcal{D}\lambda\mathcal{D}z$$
$$\cdot\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.217}$$
$$= -(-1)^{|\mathfrak{D}|}\frac{\sigma_w^2\sigma_x^2\theta_3^2}{n}\int\underbrace{\left(\prod_{(\alpha\beta)\in\mathfrak{D}}^{|\mathfrak{D}|}\frac{\sigma_w\sigma_x}{\sqrt{n}}\lambda_{\alpha\beta}\phi(z_{\alpha\beta})e^{-\sum_{\alpha\beta\in\mathfrak{D}}\left(\frac{\sigma_x^2\sigma_w^2}{2}\lambda_{\alpha\beta}^2-i\lambda_{\alpha\beta}z_{\alpha\beta}\right)}\right)}_{=(-\theta_3/n)^{|\mathfrak{D}|}}$$
$$\left(\prod_{i=1}^{n_{ind}}(\prod_{\alpha\beta\in\mathfrak{C}_i}n(-1)^{|\mathfrak{C}_i|}\underbrace{\phi(z_{\alpha\beta})\frac{\sigma_x\sigma_w}{\sqrt{n}}\lambda_{\alpha\beta})exp\left\{-\sum_{\alpha\beta\in\mathfrak{C}_i}\left(\frac{\sigma_x^2\sigma_w^2}{2}\lambda_{\alpha\beta}^2-i\lambda_{\alpha\beta}z_{\alpha\beta}\right)\right\}}_{=(-\theta_2/n)^{|\mathfrak{C}_i|}}\right)\mathcal{D}\lambda\mathcal{D}z$$
$$\cdot\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.218}$$
$$= \frac{-\sigma_w^2\sigma_x^2\theta_3^2}{n}n^{n_{ind}}\left(\frac{\theta_2}{n}\right)^{|\mathfrak{D}|+\sum_{i=1}^{n_{ind}}|\mathfrak{C}_i|}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.219}$$
$$= \frac{-\sigma_w^2\sigma_x^2\theta_3^2}{n}\left(\frac{\theta_2}{n}\right)^{|\mathfrak{D}|}\prod_{i=1}^{n_{ind}}E_{2|\mathfrak{C}_i|}^{(n)}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.220}$$

where the identities in (A.218) are taken from the study of the $2k$ cycles in (A.187)-(A.197), and $E_{2|\mathfrak{C}_i|}^{(n)}$ is the contribution of a cycle of dimension $2|\mathfrak{C}_i|$. Note that the last step is justified since if $|\mathfrak{C}_i| = 1$ we would have retrieved θ_1 rather than $n^{1-|\mathfrak{C}_i|}\theta_2^{|\mathfrak{C}_i|}$.

If instead we considered one of the expansion terms of the covariance also including one of the sets $\mathfrak{C}_i$, then the relative variables $\lambda_{\alpha\beta}$ would already appear in line (A.214), and the contribution in the log-determinant would not be relevant. Consequently the n term resulting from the expansion of the exponential of the log-determinant would not appear and we would have a relative contribution of $\mathcal{O}(1/n)$.

Therefore the contribution if the first case is

$$_{I}E^{(n)}_{\Pi_j^{(l,p)}} = -n^{n_{ind}} \frac{\sigma_w^2 \sigma_x^2 \theta_3^2}{n} \left(\frac{\theta_2}{n}\right)^l \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.221}$$

and crucially it shows that the contribution of a cycle with no factors of the type $\sigma_x \theta_3 \mathbf{W}_{i_1 p}$ can be considered independently from the block.

Second Kind: Since the block of dependent variables might contain multiple weights in the form of $\sigma_x \theta_3 \mathbf{W}_{i_1 p}$ we have to first identify the arches within the block, i.e. sequences of successive i_ξ that are all within the block. With this purpose, we consider the set of indices that define the beginning and the end of each arch

$$\mathfrak{A} = \{(i_\xi, i_{\xi+l}) \mid \exists \{\mu_j\}_{j=1}^{l} s.t. \cup_{j=1}^{l} \left((i_{\xi+j-1}, \mu_j) \cup (i_{\xi+j}, \mu_j)\right) \subset \Pi_j\}. \tag{A.222}$$

This allows to determine the contribution of $\Pi_j^{(l,\mathbf{p})}$ where l is the number of factors $\phi(\sum_l \mathbf{W}_{i_\xi,l} \mathbf{X}_{l,\mu_\xi}) \phi(\sum_l \mathbf{W}_{i_{\xi+1},l} \mathbf{X}_{l,\mu_\xi})$ in Π_j and $\mathbf{p}$ is a vector containing the indices of the columns of $\mathbf{W}$ that are considered in each cycle that Π_j intersects, and therefore has dimension $|\mathfrak{A}|$. We rewrite the expected contribution in terms of the arches which define Π_j

$$\begin{aligned} _{II}E^{(n)}_{\Pi_i^{(l,\mathbf{p})}} = \int_{\mathbf{W},\mathbf{X}} \prod_{(i_\alpha, i_{\alpha+l}) \in \mathfrak{A}} \big(-\sigma_x^2 \theta_3^2 \mathbf{W}_{i_\alpha p_\alpha} \\ \cdot \prod_{\xi=1}^{l} \left(\phi(\sum_l \mathbf{W}_{i_\xi,l} \mathbf{X}_{l,\mu_\xi}) \phi(\sum_l \mathbf{W}_{i_{\xi+1},l} \mathbf{X}_{l,\mu_\xi})\right) \mathbf{W}_{i_{\alpha+l+1} q_\alpha}\right) \mathcal{D}\mathbf{W} \mathcal{D}\mathbf{X}. \end{aligned} \tag{A.223}$$

where p_α and q_α are the dependent on the cycle in which $\mathbf{W}_{i_\alpha p}$ is. By integrating over $\mathbf{X}$ and $\mathbf{W}$ we implement similar steps as in Lemma A.2.1 were taken to to determine (A.176), we find that

$$_{II}E^{(n)}_{\Pi_i^{(l,\mathbf{p})}} = (-\sigma_x^2 \theta_3^2)^{|\mathfrak{A}|} \int \mathbb{E}\left[\prod_{(i_\alpha, i_{\alpha+l}) \in \mathfrak{A}} \mathbf{w}_{p_\alpha}^{(i_\alpha)} \mathbf{w}_{q_\alpha}^{(i_\alpha+l)}\right] \frac{1}{det(\mathbf{I} + \frac{\sigma_x^2 \sigma_w^2}{n} \mathbf{\Lambda}\mathbf{\Lambda}^\top)^{n/2}} e^{tr\mathbf{\Lambda}^\top \mathbf{Z}} F(z) \mathcal{D}\lambda \mathcal{D}z. \tag{A.224}$$

The weight vectors $\mathbf{w}_{p_\alpha}$ are distributed according to a Gaussian, therefore we can rely on Isserlis' theorem [43]. If $(X_1, \dots, X_n)$ is a zero-mean multivariate normal random vector, then

$$\mathbb{E}\left[X_1 X_2 \cdots X_n\right] = \sum_{p \in P_n^2} \prod_{\{i,j\} \in p} \mathbb{E}\left[X_i X_j\right] = \sum_{p \in P_n^2} \prod_{\{i,j\} \in p} Cov\left(X_i, X_j\right) \tag{A.225}$$

where the sum is over all the pairings of $\{1, \dots, n\}$. For example, when we consider four variables then we get that

$$\begin{aligned}\mathbb{E}\left[X_1 X_2 X_3 X_4\right] &= Cov\left(X_1, X_2\right) Cov\left(X_3, X_4\right) \\ &+ Cov\left(X_1, X_3\right) Cov\left(X_2, X_4\right) + Cov\left(X_1, X_4\right) Cov\left(X_2, X_3\right).\end{aligned} \tag{A.226}$$

We are going to consider the term in the expansion (A.225) where the covariances are only between elements in the same cycle, i.e. p_α is the same within the same cycle. The other terms have a lower order contribution to the moments because they imply two separate cycles being relative to the same sample p_α, and this implies that total contribution incurs in a cost of at least $1/n_0$.

Therefore, let's compute the contribution by defining the set of all the indices in a cycle as $\mathfrak{C}_i$, then

$${}_{II}E^{(n)}_{\Pi_i^{(l,\mathbf{p})}} = (-\sigma_x^2\theta_3^2)^{|\mathfrak{A}|} \int \mathbb{E}\left[\prod_{(i_\alpha, i_{\alpha+l}) \in \mathfrak{A}} \mathbf{w}_{p_\alpha}^{(i_\alpha)} \mathbf{w}_{q_\alpha}^{(i_\alpha+l)}\right] \frac{1}{det(\mathbf{I} + \frac{\sigma_x^2\sigma_w^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top)^{n/2}} e^{tr\mathbf{\Lambda}^\top \mathbf{Z}} F(z) \mathcal{D}\lambda \mathcal{D}z \tag{A.227}$$

$$= (-\sigma_x^2\theta_3^2)^{|\mathfrak{A}|} \int \left(\prod_{i=1}^{n_{int}} Cov\left(\mathbf{w}_{p_\alpha}^{i_\alpha} \mathbf{w}_{p_\alpha}^{i_\beta}\right)\right)\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \frac{1}{det(\mathbf{I} + \frac{\sigma_x^2\sigma_w^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top)^{n/2}} e^{tr\mathbf{\Lambda}^\top \mathbf{Z}} F(z) \mathcal{D}\lambda \mathcal{D}z \tag{A.228}$$

$$= (-\sigma_x^2\theta_3^2)^{|\mathfrak{A}|} \int \left(\prod_{i=1}^{n_{int}} Cov\left(\mathbf{w}_{p_\alpha}^{i_\alpha} \mathbf{w}_{p_\alpha}^{i_\beta}\right)\right)\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) e^{-\frac{\sigma_x^2\sigma_w^2}{2} tr(\mathbf{\Lambda}\mathbf{\Lambda}^\top) - \frac{n}{2}\sum_{\xi \geq 2} \frac{(-1)^{\xi+1}}{\xi} tr\left(\frac{\sigma_x^2\sigma_w^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top\right)^\xi - itr\mathbf{\Lambda}^\top \mathbf{Z}} F(z) \mathcal{D}\lambda \mathcal{D}z \tag{A.229}$$

In the same way as in Preposition A.1.4, the leading order of the contribution ${}_{II}E^{(n)}_{\Pi_i^{(l,\mathbf{p})}}$ is defined with a relative error $\mathcal{O}(\frac{1}{n})$ by considering the different cycles sep-

arately. Moreover, as for Lemma A.2.1 there is a relative error of $\mathcal{O}(1/n)$ by considering only the zeroth order approximation of the exponent $e^{\frac{n}{2}\sum_{\xi\geq 2}\frac{(-1)^{\xi+1}}{\xi}tr\left(\frac{\sigma_x^2\sigma_w^2}{n}\mathbf{\Lambda}\mathbf{\Lambda}^\top\right)^\xi}$. Therefore

$$
\begin{aligned}
{}_{II}E^{(n)}_{\Pi_i^{(l,\mathbf{p})}} &= (-\sigma_x^2\theta_3^2)^{|\mathfrak{A}|}\int\left(\prod_{i=1}^{n_{int}} Cov\left(\mathbf{w}_{p_\alpha}^{i_\alpha}\mathbf{w}_{p_\alpha}^{i_\beta}\right)\right)\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \\
&\quad e^{-\sum_{i=1}^{n_{int}}\left(\sum_{\alpha\beta\in\mathfrak{C}_i}(\frac{\sigma_x^2\sigma_w^2}{2}\lambda_{\alpha\beta}^2+i\lambda_{\alpha\beta}z_{\alpha\beta})\right)}F(z)\mathcal{D}\lambda\mathcal{D}z\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \qquad \text{(A.230)} \\
&= (-\sigma_x^2\theta_3^2)^{|\mathfrak{A}|}\prod_{i=1}^{n_{int}}\int\left(Cov\left(\mathbf{w}_{p_\alpha}^{i_\alpha}\mathbf{w}_{p_\alpha}^{i_\beta}\right)\right) \\
&\quad e^{-\sum_{\alpha\beta\in\mathfrak{C}_i}(\frac{\sigma_x^2\sigma_w^2}{2}\lambda_{\alpha\beta}^2+i\lambda_{\alpha\beta}z_{\alpha\beta})}F(z)\mathcal{D}\lambda\mathcal{D}z\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \qquad \text{(A.231)} \\
&= (-\sigma_x^2\theta_3^2)^{|\mathfrak{A}|}\prod_{i=1}^{n_{int}}\int\frac{\sigma_w^2}{n}\left(\prod_{\alpha\beta\in\mathfrak{C}_i}\lambda_{\alpha\beta}\phi(z_{\alpha\beta})\right) \\
&\quad e^{-\sum_{\alpha\beta\in\mathfrak{C}_i}(\frac{\sigma_x^2\sigma_w^2}{2}\lambda_{\alpha\beta}^2+i\lambda_{\alpha\beta}z_{\alpha\beta})}\mathcal{D}\lambda\mathcal{D}z\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \qquad \text{(A.232)} \\
&= (-\frac{\sigma_x^2\theta_3^2\sigma_w^2}{n})^{|\mathfrak{A}|}\left(\frac{\theta_2}{n}\right)^l\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \qquad \text{(A.233)}
\end{aligned}
$$

Note that if there is a complete cycle without factors of the type $\sigma_x\theta_3\mathbf{W}_{i_{l+1}p}$ within Π_j, then we can rely on the first case to separate its contribution obtaining a further term n_0.

Third Kind: This case is uniquely due to μ-identifications. In this case, if we have an identification at μ_ξ with $(i_\xi, i_{\xi+1}) \in \mathfrak{A}$ then we have to keep the same sample p_α in both the cycles in which $\mathfrak{A}$ is defined. Therefore

$$
{}_{III}E^{(n)}_{\Pi_i^{(l,\mathbf{p})}} = (-\frac{\sigma_x^2\theta_3^2\sigma_w^2}{n})^{|\mathfrak{A}|}\left(\frac{\theta_2}{n}\right)^l\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)\mathbb{1}_{\mathbf{p}=p\mathbf{1}} \qquad \text{(A.234)}
$$

where

$$
\mathbb{1}_{\mathbf{p}=p\mathbf{1}} = \begin{cases} 1 & \text{if } \mathbf{p} = p\mathbf{1} \text{ with } p \in [n_0] \\ 0 & \text{otherwise.} \end{cases} \qquad \text{(A.235)}
$$

and $\mathbf{1}$ is a vector whose entries are all 1.

Combination of the Contributions Let's now consider how the contribution of each identification affects the contribution of one term ω of the expansion of (A.208). For simplicity let's start by considering the case where there are no 2-dimensional cycles in the graph G, and denote with $\mathbf{p} = (p_1, \dots, p_c)$ the column index of $\mathbf{W}$ for each of the c cycle, then

$$E_{\omega}^{(n,\mathbf{p})} = \prod_{j=1}^{n_{groups}} \left(type(j) E^{(n)}_{\Pi_j^{(l_j,p_j)}} \mathbb{1}_{p=\mathbf{p}_{\alpha(j)}} \right) \left(1 + \mathcal{O}\left(\frac{1}{n}\right) \right) \tag{A.236}$$

$$= \left(\prod_{j=1}^{n_{groups}} \left(n_0^{n_{ind}^{(j)}} (-\frac{\sigma_x^2 \theta_3^2 \sigma_w^2}{n})^{|\mathfrak{A}_j|} \left(\frac{\theta_2}{n}\right)^{l_j} \mathbb{1}_{p=\mathbf{p}_{\alpha(j)}} \right) \right) \left(\prod_{\substack{(\alpha,\beta)\in \{\text{cycles connected} \\ \text{by third kind} \\ \text{identifications}\}}} \mathbb{1}_{\mathbf{p}_\alpha=\mathbf{p}_\beta} \right)$$

$$\cdot \left(1 + \mathcal{O}\left(\frac{1}{n}\right) \right) \tag{A.237}$$

where the functions $type(\cdot) : [n_{groups}] \to \{0, I, II, III\}$ and $\alpha(\cdot) : [n_{groups}] \to [n_0]$ assign to each block Π_j the typology of identification in it and the index of the column of $\mathbf{W}$ considered, and $\mathfrak{A}_j$ identifies the set of arches within a cycle for each block Π_j. This suggests that when considering the contribution of a term ω of a graph with no 2-dimensional cycles we need to keep track of the third kind of identifications since they are the only case requiring two adjacent cycles to have the same index p_α.

When considering also cycles of dimension 2, we have to distinguish between 3 cases if they are connected with a μ-identification: if the two edges within the 2-dimensional cycle are of the $\phi(\sum_l \mathbf{W}_{i_\xi,l}\mathbf{X}_{l,\mu_\xi})\phi(\sum_l \mathbf{W}_{i_{\xi+1},l}\mathbf{X}_{l,\mu_\xi})$ type, if they are of the $\sigma_x^2\theta_3^2\mathbf{W}_{i_\xi p}\mathbf{W}_{i_{\xi+1}q}$ type, or if they are alternate. In the first case, their contribution can be seen as a simple cycle and therefore we gain a factor θ_1/n rather than $\left(\frac{\theta_2}{n}\right)^l$ in equation A.237. In the second case we consider the 2 cycle as a simple block Π_j of dimension one with contribution $(-\frac{\sigma_x^2\theta_3^2\sigma_w^2}{n})$, but since this is a third case identification we need to ensure that the index p is the same between the two adjacent cycles. Finally, in the third case, we are also considering a third case identification where the intra-dependent variables of the block are the $\sigma_x\theta_3\mathbf{W}_{i_\xi p}$ edges in the 2-dimensional cycle and in the adjacent one, and all of the $\phi(\sum_l \mathbf{W}_{i_\xi,l}\mathbf{X}_{l,\mu_\xi})\phi(\sum_l \mathbf{W}_{i_{\xi+1},l}\mathbf{X}_{l,\mu_\xi})$ edges in-between them. This implies that we

gain a factor $(-\frac{\sigma_x^2\theta_3^2\sigma_w^2}{n})\left(\frac{\theta_2}{n}\right)^l\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)$ with the constraining that the index of the weight is the same as in the adjacent cycle. Note that in this case we do not have any θ_1 contribution although there is a 2-dimensional cycle.

From the above statements, it follows that there is no need to keep track of all the independent blocks Π_j to compute $E_\omega^{(n,\mathbf{p})}$; it is sufficient to identify the terms $\sigma_x^2\theta_3^2\mathbf{W}_{i_\xi p}\mathbf{W}_{i_{\xi+1}q}$, the third kind identifications, and the structure of 2-dimensional cycles. To keep track of all of these cases we will consider the following notation for each graph. We define b^μ as the number μ-identifications where at least one of the two joined cycles is of dimension 2, and c^μ as the remaining μ-identifications, i.e. $I_\mu = c^\mu + b^\mu$. We then introduce the variables $b_w^\mu \in \{0,\dots,b^\mu\}$ and $c_w^\mu \in \{0,\dots,c_w^\mu\}$ which determine how many of the b^μ and c^μ identifications have at least two edges of the $\sigma_x\theta_3\mathbf{W}_{i_\xi p}$ kind connected to the identification and therefore are of the third kind. Finally, we introduce the variables $b_w^{1/2} \in \{0,\dots,b_w^\mu\}$ and $c_w^{1/2} \in \{0,\dots,c_w^\mu\}$ which indicate if the remaining two edges are of the $\sigma_x\theta_3\mathbf{W}_{i_\xi p}$ type.

Therefore, an addend ω of a graph with $2k$ edges defined by $b^i,\ b_w^i,\ b^\mu,\ b_w^\mu,\ b_w^{1/2},\ c^i,\ c_w^i,\ c^\mu,\ c_w^\mu,\ c_w^{1/2}$ generates the following contribution

$$
\begin{aligned}
&E_\omega^{(k,\mathbf{p},b^i,b_w^i,b^\mu,b_w^\mu,b_w^{1/2},c^i,c_w^i,c^\mu,c_w^\mu,c_w^{1/2})} = \\
&= \underbrace{\underbrace{\left(\frac{\theta_1}{n}\right)^{k-b^i-2b^\mu-2c^\mu-(n_w-b_w^\mu-b_w^{1/2}-c_w^\mu-c_w^{1/2})}}_{\text{Edges of the } \phi\phi \text{ type}}\underbrace{\left(\frac{-\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{(n_w-b_w^\mu-b_w^{1/2}-c_w^\mu-c_w^{1/2})}}_{\text{Edges of the ww type}}}_{\text{Edges not adjacent to } \mu\text{-identification and in cycle larger than 2}} \\
&\cdot \underbrace{\underbrace{\left(\frac{\theta_2}{n}\right)^{b_w^\mu-b_w^{1/2}}\left(\frac{-\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{b_w^\mu-b_w^{1/2}}}_{\text{Mixed Case}}\underbrace{\left(\frac{-\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{2b_w^{1/2}}}_{\text{Edges of the } ww \text{ type}}\underbrace{\left(\frac{\theta_1}{n}\right)^{b^\mu-b_w^\mu}\left(\frac{-\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{b^\mu-b_w^\mu}}_{\text{Edges of the } \phi\phi \text{ type}}}_{\text{Edges adjacent to } \mu\text{-identification with a cycle of dimension 2}} \\
&\cdot \underbrace{\underbrace{\left(\frac{\theta_2}{n}\right)^{c_w^\mu-c_w^{1/2}}\left(\frac{-\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{c_w^\mu-c_w^{1/2}}}_{\text{Mixed Case}}\underbrace{\left(\frac{-\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{2c_w^{1/2}}}_{\text{Edges of the } ww \text{ type}}\underbrace{\left(\frac{-\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{2(c^\mu-c_w^\mu)}}_{\text{Edges of the } \phi\phi \text{ type}}}_{\text{Edges adjacent to } \mu\text{-identification and in cycles larger than 2}}
\end{aligned}
$$

$$\cdot \underbrace{\left(\frac{\theta_1}{n}\right)^{b^i-b_w^i}\left(\frac{-\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{b_w^i}}_{\substack{\text{Adjacent to i-identification}\\ \text{and with a cycle of dimension 2}}}\left(\prod_{\substack{(\alpha,\beta)\in \{\text{cycles connected}\\ \text{by third kind}\\ \text{identifications}\}}}\mathbb{1}_{\mathbf{p}_\alpha=\mathbf{p}_\beta}\right)\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right). \tag{A.238}$$

For each of the addends ω, there are $n_0^{1+I_\mu+I_i}$ ways to choose the indices p for each of the $1+I_\mu+I_i$ cycles. However, for all the identifications of third kind there is a constraint on adjacent cycles having the same index p, therefore there are actually $1+I_\mu+I_i-b_w^\mu-c_w^\mu$ non zero terms ω sharing the same structure $(k,\cdot,b^i,b_w^i,b^\mu,b_w^\mu,b_w^{1/2},c^i,c_w^i,c^\mu,c_w^\mu,c_w^{1/2})$ and therefore having the same contribution.

Now let's consider the contribution of the full graph by considering all of the combinations on how the structure $(k,\mathbf{p},b^i,b_w^i,b^\mu,b_w^\mu,b_w^{1/2},c^i,c_w^i,c^\mu,c_w^\mu,c_w^{1/2})$ may change. Let's start by considering the edges located in any of the $\sum_i k_i-b^\mu-2^{c^\mu}$ pair of edges that are not adjacent to a μ-identification and are in cycles bigger than 2, there are $\binom{k-b^i-2b^\mu-2^{c^\mu}}{l}$ ways for l terms $\sigma_x^2\theta_3^2\mathbf{W}_{i_\xi p}\mathbf{W}_{i_{\xi+1}q}$ to be located. Similarly, there are $\binom{b^i}{b_w^i}$ ways to have b_w^i 2-dimensional cycles whose edges are of the $\sigma_x^2\theta_3^2\mathbf{W}_{i_\xi p}\mathbf{W}_{i_{\xi+1}q}$ kind. Considering the I_μ identifications, there are $\binom{b^\mu}{b_w^\mu}$ and $\binom{c^\mu}{c_w^\mu}$ ways to choose only one pair in the identification to be of the $\sigma_x^2\theta_3^2\mathbf{W}_{i_\xi p}\mathbf{W}_{i_{\xi+1}q}$ type for the case with 2-dimensional cycles or higher dimensions. Finally if a μ-identification is such that all the four edges are of the $\sigma_x^2\theta_3^2\mathbf{W}_{i_\xi p}\mathbf{W}_{i_{\xi+1}q}$ kind, there are $\binom{b^\mu}{b_w^{1/2}}$ and $\binom{c^\mu}{c_w^{1/2}}$ choices in the $b_w^{1/2}$ and $c_w^{1/2}$ cases.

Combining all of the contributions above, it is possible to compute the contribution of a graph G with $2k$ edges and with I_μ and I_i identifications, of which b^μ and b^i generate a 2-dimensional cycle.

$$\begin{aligned} E_{\mathcal{G}} &= \sum_{l}^{k-b^i-2b^\mu-2c^\mu}\sum_{c_w^\mu}^{c^\mu}\sum_{c_w^{1/2}}^{c_w^\mu}\sum_{b_w^\mu}^{b^\mu}\sum_{b_w^{1/2}}^{b_w^\mu}\sum_{b_w^i}^{b^i}\binom{k-b^i-2b^\mu-2c^\mu}{l}\binom{c^\mu}{c_w^\mu}\binom{c_w^\mu}{c_w^{1/2}}\binom{b^\mu}{b_w^\mu}\binom{b_w^\mu}{b_w^{1/2}}\binom{b^i}{b_w^i}\\ &\quad\cdot E_\omega^{(k,\mathbf{p},b^i,b_w^i,b^\mu,b_w^\mu,b_w^{1/2},c^i,c_w^i,c^\mu,c_w^\mu,c_w^{1/2})}n_0^{1+I_\mu+I_i-b_w^\mu-c_w^\mu} \qquad\qquad (A.239)\\ &= \sum_{l}^{k-b^i-2b^\mu-2c^\mu}\sum_{c_w^\mu}^{c^\mu}\sum_{c_w^{1/2}}^{c_w^\mu}\sum_{b_w^\mu}^{b^\mu}\sum_{b_w^{1/2}}^{b_w^\mu}\sum_{b_w^i}^{b^i}\binom{k-b^i-2b^\mu-2c^\mu}{l}\left(\frac{\theta_2}{n}\right)^{k-b^i-2b^\mu-2c^\mu-l}\left(-\frac{\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^l \end{aligned}$$

$$\binom{c^\mu}{c_w^\mu}\left(\frac{\theta_2}{n}\right)^{c^\mu-c_w^\mu}\left(-\frac{\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{c_w^\mu}\binom{c_w^\mu}{c_w^{1/2}}\left(\frac{\theta_2}{n}\right)^{c^\mu-c_w^{1/2}}\left(-\frac{\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{c_w^{1/2}}$$
$$\binom{b^\mu}{b_w^\mu}\left(\frac{\theta_1}{n}\right)^{b^\mu-b_w^\mu}\left(-\frac{\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{b_w^\mu}\binom{b_w^\mu}{b_w^{1/2}}\left(\frac{\theta_2}{n}\right)^{b^\mu-b_w^{1/2}}\left(-\frac{\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{b_w^{1/2}}$$
$$\binom{b^i}{b_w^i}\left(\frac{\theta_1}{n}\right)^{b^i-b_w^i}\left(-\frac{\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{b_w^i} n_0^{1+I_\mu+I_i-b_w^\mu-c_w^\mu}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.240}$$

Let's now group all the factors that are independent from each other and consider their sum and consider $\left(\frac{\theta_2}{n}\right)^{c^\mu-c_w^{1/2}} = \left(\frac{\theta_2}{n}\right)^{c^\mu-c_w^\mu}\left(\frac{\theta_2}{n}\right)^{c_w^\mu-c_w^{1/2}}$ and $\left(\frac{\theta_2}{n}\right)^{b^\mu-b_w^{1/2}} = \left(\frac{\theta_2}{n}\right)^{b^\mu-b_w^\mu}\left(\frac{\theta_2}{n}\right)^{b_w^\mu-b_w^{1/2}}$.

$$\begin{aligned}
E_{\mathcal{G}} = n_0^{1+I_\mu+I_i}&\left[\sum_l^{k-b^i-2b^\mu-2c^\mu}\binom{k-b^i-2b^\mu-2c^\mu}{l}\left(\frac{\theta_2}{n}\right)^{k-b^i-2b^\mu-2c^\mu-l}\left(-\frac{\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^l\right]\\
&\cdot\left[\sum_{b_w^i}^{b^i}\binom{b^i}{b_w^i}\left(\frac{\theta_1}{n}\right)^{b^i-b_w^i}\left(-\frac{\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{b_w^i}\right]\\
&\cdot\left[\sum_{c_w^\mu}^{c^\mu}\binom{c^\mu}{c_w^\mu}\left(\frac{\theta_2^2}{n^2}\right)^{c^\mu-c_w^\mu}\left(-\frac{\sigma_x^2\sigma_w^2\theta_3^2}{n^2}\right)^{c_w^\mu}\right.\\
&\qquad\left.\cdot\left(\sum_{c_w^{1/2}}^{c_w^\mu}\binom{c_w^\mu}{c_w^{1/2}}\left(\frac{\theta_2}{n}\right)^{c_w^\mu-c_w^{1/2}}\left(\frac{-\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{c_w^{1/2}}\right)\right]\\
&\cdot\left[\sum_{b_w^\mu}^{b^\mu}\binom{b^\mu}{b_w^\mu}\left(\frac{\theta_1\theta_2}{n^2}\right)^{b^\mu-b_w^\mu}\left(-\frac{\sigma_x^2\sigma_w^2\theta_3^2}{n^2}\right)^{b_w^\mu}\right.\\
&\qquad\left.\cdot\left(\sum_{b_w^{1/2}}^{b_w^\mu}\binom{b_w^\mu}{b_w^{1/2}}\left(\frac{\theta_2}{n}\right)^{b_w^\mu-b_w^{1/2}}\left(\frac{-\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{b_w^{1/2}}\right)\right]\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right)
\end{aligned} \tag{A.241}$$

$$\begin{aligned}
&= n_0^{1+I_\mu+I_i}\left(\frac{\theta_2}{n}-\frac{\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{k-b^i-2b^\mu-2c^\mu}\left(\frac{\theta_1}{n}-\frac{\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{b^i}\\
&\quad\cdot\left[\sum_{c_w^\mu}^{c^\mu}\binom{c^\mu}{c_w^\mu}\left(\frac{\theta_2^2}{n^2}\right)^{c^\mu-c_w^\mu}\left(-\frac{\sigma_x^2\sigma_w^2\theta_3^2}{n^2}\right)^{c_w^\mu}\left(\frac{\theta_2}{n}-\frac{\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{c_w^\mu}\right]
\end{aligned}$$

$$\cdot \left[\sum_{b_w^\mu}^{b^\mu} \binom{b^\mu}{b_w^\mu} \left(\frac{\theta_1\theta_2}{n^2}\right)^{b^\mu - b_w^\mu} \left(-\frac{\sigma_x^2\sigma_w^2\theta_3^2}{n^2}\right)^{b_w^\mu} \left(\frac{\theta_2}{n} - \frac{\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{b_w^\mu}\right] \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.242}$$

$$= n_0^{1+I_\mu+I_i} \left(\frac{\theta_2}{n} - \frac{\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{k-b^i-2b^\mu-2c^\mu} \left(\frac{\theta_1}{n} - \frac{\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{b^i}$$
$$\cdot \left[\frac{\theta_2^2}{n^2} - \frac{\sigma_x^2\sigma_w^2\theta_3^2}{n^2}\left(\frac{\theta_2}{n} - \frac{\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)\right]^{c^\mu} \left[\frac{\theta_1\theta_2}{n^2} - \frac{\sigma_x^2\sigma_w^2\theta_3^2}{n^2}\left(\frac{\theta_2}{n} - \frac{\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)\right]^{b^\mu}$$
$$\cdot \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.243}$$

$$= n_0^{1+I_\mu+I_i} \left(\frac{\theta_2}{n} - \frac{\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{k-b^i-2b^\mu-2c^\mu} \left(\frac{\theta_1}{n} - \frac{\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{b^i}$$
$$\cdot \left[\left(\frac{\theta_2}{n} - \frac{\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^2 \left(1 + \frac{2\sigma_x^2\sigma_w^2\theta_3^2\theta_2 - \sigma_x^4\sigma_w^4\theta_3^4}{(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)^2} + \mathcal{O}\left(\frac{1}{n}\right)\right)\right]^{c^\mu}$$
$$\cdot \left[\left(\frac{\theta_2}{n} - \frac{\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)\left(\frac{\theta_1}{n} - \frac{\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)\left(1 + \frac{\theta_1\sigma_x^2\sigma_w^2\theta_3^2 + \sigma_x^2\sigma_w^2\theta_3^2\theta_2 - \sigma_x^4\sigma_w^4\theta_3^4}{(\theta_1 - \sigma_x^2\sigma_w^2\theta_3^2)(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)} + \mathcal{O}\left(\frac{1}{n}\right)\right)\right]^{b^\mu}$$
$$\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.244}$$

$$= n_0^{1+I_\mu+I_i} \left(\frac{\theta_2}{n} - \frac{\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{k-b^i-b^\mu} \left(\frac{\theta_1}{n} - \frac{\sigma_x^2\sigma_w^2\theta_3^2}{n}\right)^{b^i+b^\mu}$$
$$\left[1 + \frac{2\sigma_x^2\sigma_w^2\theta_3^2\theta_2 - \sigma_x^4\sigma_w^4\theta_3^4}{(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)^2} + \mathcal{O}\left(\frac{1}{n}\right)\right]^{c^\mu}$$
$$\left[1 + \frac{\theta_1\sigma_x^2\sigma_w^2\theta_3^2 + \sigma_x^2\sigma_w^2\theta_3^2\theta_2 - \sigma_x^4\sigma_w^4\theta_3^4}{(\theta_1 - \sigma_x^2\sigma_w^2\theta_3^2)(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)} + \mathcal{O}\left(\frac{1}{n}\right)\right]^{b^\mu}$$
$$\left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.245}$$

$$= n_0^{1+I_\mu+I_i-k} \left(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2\right)^{\sum_i k_i} \left(\theta_1 - \sigma_x^2\sigma_w^2\theta_3^2\right)^{b^i+b^\mu}$$
$$\left[1 + \frac{2\sigma_x^2\sigma_w^2\theta_3^2\theta_2 - \sigma_x^4\sigma_w^4\theta_3^4}{(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)^2}\right]^{c^\mu} \left[1 + \frac{\theta_1\sigma_x^2\sigma_w^2\theta_3^2 + \sigma_x^2\sigma_w^2\theta_3^2\theta_2 - \sigma_x^4\sigma_w^4\theta_3^4}{(\theta_1 - \sigma_x^2\sigma_w^2\theta_3^2)(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)}\right]^{b^\mu}$$
$$\cdot \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.246}$$

$$= n_0^{c-\sum_i k_i} \left(\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2\right)^{\sum_i k_i} \left(\theta_1 - \sigma_x^2\sigma_w^2\theta_3^2\right)^{b^i+b^\mu}$$

$$\left[1+\frac{2\sigma_x^2\sigma_w^2\theta_3^2\theta_2-\sigma_x^4\sigma_w^4\theta_3^4}{(\theta_2-\sigma_x^2\sigma_w^2\theta_3^2)^2}\right]^{c^\mu}\left[1+\frac{\theta_1\sigma_x^2\sigma_w^2\theta_3^2+\sigma_x^2\sigma_w^2\theta_3^2\theta_2-\sigma_x^4\sigma_w^4\theta_3^4}{(\theta_1-\sigma_x^2\sigma_w^2\theta_3^2)(\theta_2-\sigma_x^2\sigma_w^2\theta_3^2)}\right]^{b^\mu}$$
$$\cdot\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.247}$$

where in the last inequality we used the two equalities $k=\sum_i k_i+b^i+b^\mu$ and $1+I_\mu+I_i=c+b^i+b^\mu$. □

A.2.3.1 Proof of Proposition 3.3.4

Proof.

$$\mathfrak{m}_k^{(n)}=\frac{1}{n_1m^k}\sum_{I_i,I_\mu=1}^{k}\sum_{b=0}^{I_i+I_\mu+1}\sum_{b^\mu=0}^{b}\mathcal{C}(k,I_i,I_\mu,b,b^\mu,c^\mu)E_{\mathcal{G}(k,I_i,I_\mu,b,b^\mu,c^\mu)}\#_1\#_2 \tag{A.248}$$

$$=\frac{1}{n_1m^k}\sum_{I_i,I_\mu=1}^{k}\sum_{b=0}^{I_i+I_\mu+1}\sum_{b^\mu=0}^{b}\mathcal{C}(k,I_i,I_\mu,b,b^\mu,c^\mu)n_0^{1+I_i+I_\mu-b-(k-b)}\cdot$$
$$\cdot(\theta_2-\sigma_x^2\sigma_w^2\theta_3^2)^{k-b}(\theta_1-\sigma_x^2\sigma_w^2\theta_3^2)^b\kappa_b^{b^\mu}\kappa_c^{I_\mu-b^\mu}n_1^{k-I_1}m^{k-I_\mu}\left(1+\mathcal{O}\left(\frac{1}{n_0}\right)\right)$$
$$\cdot\left(\frac{e}{k-I_i}\right)^{k-I_1}(2\pi(k-I_i))^{-1/2}\left(\frac{e}{k-I_\mu}\right)^{k-I_\mu}(2\pi(k-I_\mu))^{-1/2}$$
$$\cdot\left(1+o\left(\frac{1}{m}\right)\right)\left(1+o\left(\frac{1}{n_1}\right)\right) \tag{A.249}$$

$$=\sum_{I_i,I_\mu=1}^{k}\sum_{b=0}^{I_i+I_\mu+1}\sum_{b^\mu=0}^{b}\mathcal{C}(k,I_i,I_\mu,b,b^\mu,c^\mu)(\theta_1-\sigma_x^2\sigma_w^2\theta_3^2)^b\cdot$$
$$\cdot(\theta_2-\sigma_x^2\sigma_w^2\theta_3^2)^{k-b}\frac{n_1^{k-1}}{n_0^{k-1}}\frac{n_0^{I_1}}{n_1^{I_1}}\frac{n_0^{I_\mu}}{m^{I_\mu}}\kappa_b^{b^\mu}\kappa_c^{I_\mu-b^\mu}$$
$$\cdot\left(\frac{e}{k-I_i}\right)^{k-I_1}\left(\frac{e}{k-I_\mu}\right)^{k-I_\mu}\frac{1}{2\pi\sqrt{(k-I_i)(k-I_\mu)}}$$
$$\cdot\left(1+\mathcal{O}\left(\frac{1}{n_0}\right)\right) \tag{A.250}$$

$$=\varphi^{1-k}\sum_{I_i,I_\mu=1}^{k}\sum_{b=0}^{I_i+I_\mu+1}\sum_{b^\mu=0}^{b}\frac{\mathcal{C}(k,I_i,I_\mu,b,b^\mu,c^\mu)}{2\pi\sqrt{(k-I_i)(k-I_\mu)}}(\theta_1-\sigma_x^2\sigma_w^2\theta_3^2)^b\cdot$$
$$\cdot(\theta_2-\sigma_x^2\sigma_w^2\theta_3^2)^{k-b}\kappa_b^{b^\mu}\kappa_c^{I_\mu-b^\mu}\varphi^{I_1}\psi^{I_\mu}$$
$$\cdot\left(\frac{e}{k-I_i}\right)^{k-I_1}\left(\frac{e}{k-I_\mu}\right)^{k-I_\mu}\left(1+\mathcal{O}\left(\frac{1}{n_0}\right)\right) \tag{A.251}$$

where we used the coefficients $\varphi = n_0/n_1$, $\psi = n_0/m$. Therefore

$$\mathfrak{m}_k = \lim_{n\to\infty} \mathfrak{m}_k^{(n)} \tag{A.252}$$

$$= \varphi^{1-k} \sum_{I_i,I_\mu=1}^{k} \sum_{b=0}^{I_i+I_\mu+1} \sum_{b^\mu=0}^{b} \frac{\mathcal{C}(k, I_i, I_\mu, b, b^\mu, c^\mu)}{2\pi\sqrt{(k-I_i)(k-I_\mu)}} (\theta_1 - \sigma_x^2\sigma_w^2\theta_3^2)^b \cdot$$

$$\cdot (\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)^{k-b} \kappa_b^{b^\mu} \kappa_c^{I_\mu - b^\mu} \varphi^{I_1} \psi^{I_\mu} \left(\frac{e}{k-I_i}\right)^{k-I_1} \left(\frac{e}{k-I_\mu}\right)^{k-I_\mu} \tag{A.253}$$

and $|\mathfrak{m}_k^{(n)} - \mathfrak{m}_k| = \mathcal{O}\left(\frac{1}{n}\right)$. □

A.2.4 Side Theorems

Convergence to the empirical spectral measure As for the case with the correlated input, the explicit Stieltjes transform in Corollary 3.3.4.1 is too computationally intensive to be used to compute the transform. However, it allows to check that the Carleman's condition in Theorem A.1.5 for the spectral measure defining the moments to be unique [87]. This is done, by ensuring that $\mathfrak{m}_k < C^k$.

Lemma A.2.3. *For* $\boldsymbol{M} = \frac{1}{m}\mathbb{E}_{\boldsymbol{X}}\left[\boldsymbol{Y}\boldsymbol{Y}^\top\right] - \frac{1}{\sigma_x^2 m^2}\mathbb{E}_{\boldsymbol{x}}\left[\boldsymbol{Y}\boldsymbol{X}^\top\right]\mathbb{E}_{\boldsymbol{X}}\left[\boldsymbol{X}\ \boldsymbol{Y}^\top\right]$ *and* $\boldsymbol{Y} = \phi(\boldsymbol{W}\boldsymbol{X})$ *as defined in Theorem 3.3.1, there exists a variable* C *such that*

$$\mathfrak{m}_k \leq C^k. \tag{A.254}$$

Proof. Following the same steps as in the proof the Lemma A.1.6 based on [6], there is a constant $\bar{C}$ such that $\mathcal{C}(k, I_i, I_\mu, b, b^\mu, c^\mu) < \bar{C}\frac{3\tau}{4\sqrt{\pi}}\frac{\nu^{k-3/2}}{k^{5/2}}$ and consequently

$$\mathfrak{m}_k = \varphi^{1-k} \sum_{I_i,I_\mu=1}^{k} \sum_{b=0}^{I_i+I_\mu+1} \sum_{b^\mu=0}^{b} \frac{\mathcal{C}(k, I_i, I_\mu, b, b^\mu, c^\mu)}{2\pi\sqrt{(k-I_i)(k-I_\mu)}} (\theta_1 - \sigma_x^2\sigma_w^2\theta_3^2)^b (\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)^{k-b}$$

$$\kappa_b^{b^\mu} \kappa_c^{I_\mu - b^\mu} \varphi^{I_1} \psi^{I_\mu} \left(\frac{e}{k-I_i}\right)^{k-I_1} \left(\frac{e}{k-I_\mu}\right)^{k-I_\mu} \tag{A.255}$$

$$\leq \varphi^{1-k} \sum_{I_i,I_\mu=1}^{k} \sum_{b=0}^{I_i+I_\mu+1} \sum_{b^\mu=0}^{b} \frac{3\bar{C}\tau}{2\pi 4\sqrt{\pi}} \frac{\nu^{k-3/2}}{k^{5/2}} (\theta_1 - \sigma_x^2\sigma_w^2\theta_3^2)^b (\theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)^{k-b}$$

$$\kappa_b^{b^\mu} \kappa_c^{I_\mu - b^\mu} \varphi^{I_1} \psi^{I_\mu} (e)^{k-I_1} (e)^{k-I_\mu} \tag{A.256}$$

$$\leq \varphi^{1-k} \sum_{I_i,I_\mu=1}^{k} \sum_{b=0}^{I_i+I_\mu+1} \sum_{b^\mu=0}^{b} \frac{3\bar{C}\tau}{2\pi 4\sqrt{\pi}} \frac{\nu^{k-3/2}}{k^{5/2}} \max(1, \theta_1 - \sigma_x^2\sigma_w^2\theta_3^2)^k \max(1, \theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)^k$$

$$\max(1, \kappa_b)^k \max(1, \kappa_c)^{I_\mu} \left(\frac{\varphi}{e}\right)^{I_1} \left(\frac{\psi}{e}\right)^{I_\mu} e^{2k} \tag{A.257}$$

$$\leq \varphi^{1-k} \frac{3\bar{C}\tau}{2\pi 4\sqrt{\pi}} \frac{\nu^{k-3/2}}{k^{5/2}} \max(1, \theta_1 - \sigma_x^2\sigma_w^2\theta_3^2)^k \max(1, \theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)^k \max(1, \kappa_b)^k e^{2k}$$

$$\sum_{I_i,I_\mu=1}^{k} \max(1, \kappa_c)^{I_\mu} \left(\frac{\varphi}{e}\right)^{I_1} \left(\frac{\psi}{e}\right)^{I_\mu} (I_i + I_\mu + 1)(I_i + I_\mu + 2)/2 \tag{A.258}$$

$$\leq \varphi^{1-k} \frac{3\bar{C}\tau}{2\pi 4\sqrt{\pi}} \frac{\nu^{k-3/2}}{k^{5/2}} \max(1, \theta_1 - \sigma_x^2\sigma_w^2\theta_3^2)^k \max(1, \theta_2 - \sigma_x^2\sigma_w^2\theta_3^2)^k \max(1, \kappa_b)^k$$

$$\max(1, \kappa_c)^k \left(\frac{\varphi}{e}\right)^{k} \left(\frac{\psi}{e}\right)^{k} (2k + 1)(2k + 2)/2 \tag{A.259}$$

$$\leq C^k \tag{A.260}$$

and therefore the Carleman's condition is satisfied and there is a unique distribution μ defined by the considered Stieltjes transform. □

Contribution of non-admissible graphs The computation of the moments relies on the set of admissible graphs, as defined in Definition 3.1.1. The contributions of the addends in (3.41) with non-admissible graphs is not significant to the moments of the eigen-distribution as in the corrleated input case. This is because the argument in [72, Supplementary Material 1.2.1] and [6, Section 3.1.4] showing that the leading order contribution for non-admissible graphs is $\mathcal{O}(n_0^{c-1-k})$ rather than $\mathcal{O}(n_0^{c-k})$ as for the admissible graphs, relies on the fact that non-admissible graphs require a further identification and this does not change in this case. This statement relies on the assumption that the activation function is such that $\left|\int \phi^k(\sigma_w\sigma_x z)\mathcal{D}z\right| < \infty$ since otherwise the contribution of a graph that consists in going through the same 2-dimensional cycle k-times could blow up. Therefore, non-admissible graphs generate a negligible contribution for $n \to \infty$.

A.2.5 Proof Corollary 3.3.1.1

Proof. Becuase of $\theta_2 = \sigma_x^2\sigma_w^2\theta_3^2$ at the first layer and of considering $\psi = 0$ and $\varphi = 1$, then we find that the recursive relation for H is

$$H(z) = 1 + \frac{H_{\psi b}(z)H_{\varphi}(z)(\theta_1 - \sigma_x^2\sigma_{\mathbf{W}}^2\theta_3^2)}{\varphi z} \tag{A.261}$$

$$= 1 + \frac{H(z)(\theta_1 - \theta_2)}{z} \tag{A.262}$$

and therefore

$$\frac{H}{z} = \frac{1}{z - (\theta_1 - \theta_2)}. \tag{A.263}$$

By expliciting the related Stieltjes transform we find that

$$G(z) = \frac{H}{z} = \frac{1}{z - (\theta_1 - \theta_2)}. \tag{A.264}$$

which implies that the covariance matrix $\frac{1}{m}\mathbb{E}_{\mathbf{X}}\left[\mathbf{Y}\mathbf{Y}^\top\right] - \frac{1}{\sigma_x^2 m^2}\mathbb{E}_{\mathbf{x}}\left[\mathbf{Y}\mathbf{X}^\top\right]\mathbb{E}_{\mathbf{X}}\left[\mathbf{X}\ \mathbf{Y}^\top\right]$ is the identity matrix scaled by $(\theta_1 - \theta_2)$. Therefore, the distribution of the post-activation layer is going to be defined by $\mathbf{WW}$. □

A.3 Proofs Relevant to Theorem 3.3.5

In this section, it is going to be necessary computing expectations over the variable

$$\mathfrak{W}_{\alpha\beta}^{(\ell)} = \sum_{k_\ell} \mathbf{W}_{\alpha k_\ell}^{(\ell)} \left(\prod_{j=1}^{\ell-2} \left(\sum_{k_{\ell-j}} \mathbf{W}_{k_{\ell-j+1}k_{\ell-j}}^{(\ell-j)}) \right) \right) \mathbf{W}_{k_2\beta}^{(1)}.$$

Specifically the following Lemma is going to be used.

Lemma A.3.1. *Consider a succession of independent matrices* $\{\boldsymbol{W}^{(l)}\}_{l=1}^{\ell}$ *such that* $\boldsymbol{W}^{(l)} \in \mathbb{R}^{n_l \times n_{l-1}}$ *and define the following variable*

$$\mathfrak{W}_{\alpha\beta}^{(\ell)} = \sum_{k_\ell} \boldsymbol{W}_{\alpha k_\ell}^{(\ell)} \left(\prod_{j=1}^{\ell-2} \left(\sum_{k_{\ell-j}} \boldsymbol{W}_{k_{\ell-j+1}k_{\ell-j}}^{(\ell-j)}) \right) \right) \boldsymbol{W}_{k_2\beta}^{(1)}.$$

Then

$$\int \mathfrak{W}_{\alpha\beta}^{(\ell)}\mathfrak{W}_{\gamma\beta}^{(\ell)}\mathcal{D}\mathfrak{W}^{(\ell)} = \mathbb{1}_{\{\gamma\}}(\alpha)\frac{\sigma_w^{2\ell}}{n} \tag{A.265}$$

where $\mathbb{1}_{\{\gamma\}}$ *is the indicator function, i.e.* $\mathbb{1}_{\{\gamma\}}(\alpha) = 1$ *if* $\alpha = \gamma$ *and* $\mathbb{1}_{\{\gamma\}}(\alpha) = 0$ *if* $\alpha \neq \gamma$.

Proof. Consider $\alpha \neq \gamma$, then

$$\int \mathfrak{W}_{\alpha\beta}^{(\ell)} \mathfrak{W}_{\gamma\beta}^{(\ell)} \mathcal{D}\mathfrak{W}^{(\ell)} = \int \sum_{k_1,k_2=1}^{n} \mathbf{W}_{\alpha k_1}^{(\ell)} \mathfrak{W}_{k_1\beta}^{(\ell-1)} \mathbf{W}_{\gamma k_2}^{(\ell)} \mathfrak{W}_{k_2\beta}^{(\ell-1)} \mathcal{D}\mathbf{W}^{(\ell)} \mathcal{D}\mathfrak{W}^{(\ell-1)} \tag{A.266}$$

$$= \sum_{k_1,k_2=1}^{n} \underbrace{\int \mathbf{W}_{\alpha k_1}^{(\ell)} \mathbf{W}_{\gamma k_2}^{(\ell)} \mathcal{D}\mathbf{W}^{(\ell)}}_{=0} \int \mathfrak{W}_{k_1\beta}^{(\ell-1)} \mathfrak{W}_{k_2\beta}^{(\ell-1)} \mathcal{D}\mathfrak{W}^{(\ell-1)} = 0. \tag{A.267}$$

While, if $\alpha = \gamma$

$$\int {\mathfrak{W}_{\alpha\beta}^{(\ell)}}^2 \mathcal{D}\mathfrak{W}^{(\ell)} = \int \sum_{k_1,k_2,\ldots,k_{\ell-1}=1}^{n} \left(\mathbf{W}_{\alpha k_1}^{(\ell)} \mathbf{W}_{k_1 k_2}^{(\ell-1)} \ldots \mathbf{W}_{k_{\ell-1}\beta}^{(1)} \right)^2 \mathcal{D}\mathbf{W}^{(\ell)} \mathcal{D}\mathbf{W}^{(\ell-1)} \ldots \mathcal{D}\mathbf{W}^{(1)} \tag{A.268}$$

$$= \sum_{k_1,k_2,\ldots,k_{\ell-1}=1}^{n} \left(\int {\mathbf{W}_{\alpha k_1}^{(\ell)}}^2 \mathcal{D}\mathbf{W}^{(\ell)} \int {\mathbf{W}_{k_1 k_2}^{(\ell-1)}}^2 \mathcal{D}\mathbf{W}^{(\ell-1)} \ldots \int {\mathbf{W}_{k_{\ell-1}\beta}^{(1)}}^2 \mathcal{D}\mathbf{W}^{(1)} \right) \tag{A.269}$$

$$= \sum_{k_1,k_2,\ldots,k_{\ell-1}=1}^{n} \left(\frac{\sigma_w^2}{n} \right)^{\ell} = n^{\ell-1} \left(\frac{\sigma_w^2}{n} \right)^{\ell} = \frac{\sigma_w^{2\ell}}{n} \tag{A.270}$$

where we used in (A.268) the i.i.d. nature of the terms in $\mathfrak{W}_{\alpha\beta}$ for whom the following statement holds: if two variables X_1 and X_2 are independent then $\mathbb{E}[(X_1 + X_2)^2] = \mathbb{E}[X_1^2] + \mathbb{E}[X_2^2]$ since $Cov[X_1 X_2] = 0$. □

A.3.1 Proof of Lemma 3.3.6

Proof. In the case where $k = 1$ we then find

$$E_2^{(n,\ell)} = \int_{\mathbf{W},\mathbf{Y}} \phi(\sum_l \mathbf{W}_{i_\xi,l}^{(\ell)} \mathbf{Y}_{l,\mu_\xi}^{(\ell)}) \phi(\sum_l \mathbf{W}_{i_\xi,l}^{(\ell)} \mathbf{Y}_{l,\mu_\xi}^{(\ell)}) - \sigma_x^2 \sum_{p=1}^{n} \mathfrak{W}_{i_\xi p}^{(\ell)} \mathfrak{W}_{i_\xi p}^{(\ell)} \left(\prod_{j=1}^{\ell} \phi'(\mathbf{H}_{k_\ell p}^{(\ell)}) \right) \mathcal{D}\mathbf{W} \mathcal{D}\mathbf{X} \tag{A.271}$$

$$= \int \tilde{\theta}_1^{(n,\ell)} \mathcal{D}\mathbf{\Sigma}^{(\ell)} \left(1 + \mathcal{O}\left(\frac{1}{n} \right) \right) - \sigma_x^2 \prod_{l=1}^{\ell} (\theta_3^{(\ell)})^2 \int_{\mathbf{W}} \sum_{p=1}^{n} \mathfrak{W}_{i_\xi p}^{(\ell)} \mathfrak{W}_{i_\xi p}^{(\ell)} \mathcal{D}\mathbf{W} \tag{A.272}$$

$$= \int \tilde{\theta}_1^{(n,\ell)} \mathcal{D}\mathbf{\Sigma}^{(\ell)} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) - \sigma_x^2 \prod_{l=1}^{\ell} (\theta_3^{(\ell)})^2 \sum_{p=1}^{n} \frac{\sigma_w^{2\ell}}{n} \tag{A.273}$$

$$= \int \tilde{\theta}_1^{(n,\ell)} \mathcal{D}\mathbf{\Sigma}^{(\ell)} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right) - \sigma_x^2 \sigma_w^{2\ell} \prod_{l=1}^{\ell} (\theta_3^{(\ell)})^2 \tag{A.274}$$

where we followed the same process as in Lemma 3.2.3 to compute the integral

$$\int_{\mathbf{W},\mathbf{Y}} \phi(\sum_l \mathbf{W}_{i_\xi,l}^{(\ell)} \mathbf{Y}_{l,\mu_\xi}^{(\ell)}) \phi(\sum_l \mathbf{W}_{i_\xi,l}^{(\ell)} \mathbf{Y}_{l,\mu_\xi}^{(\ell)}) \mathcal{D}\mathbf{W}\mathcal{D}\mathbf{X} = \int \tilde{\theta}_1^{(n,\ell)} \mathcal{D}\mathbf{\Sigma}^{(\ell)} \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right)$$

with

$$\tilde{\theta}_1^{(n,\ell)} = \frac{1}{\sqrt{2\pi}} \phi(\sqrt{q^{(\ell)}} \sqrt{\frac{tr(\mathbf{\Sigma}^{(\ell)})}{n}} z)^2 e^{-\frac{z^2}{2}} dz \tag{A.275}$$

$$\theta_3^{(\ell)} = \int \phi'(\sqrt{q^{(\ell)}} z) e^{-\frac{z^2}{2}} dz. \tag{A.276}$$

□

A.3.2 Lemma 3.3.7

When considering $k > 1$ it is necessary to consider all the mixed products. As a matter of fact each term in the expansion of the product in $E_{2k}^{(n,\ell)}$ consists of successions of products of the kind $\phi(\sum_l \mathbf{W}_{i_\xi,l}^{(\ell)} \mathbf{Y}_{l,\mu_\xi}^{(\ell)}) \phi(\sum_l \mathbf{W}_{i_{\xi+1},l}^{(\ell)} \mathbf{Y}_{l,\mu_\xi}^{(\ell)})$ alternated with products of the kind $\sigma_x^2 \sum_{p=1}^{n} \mathfrak{W}_{i_\xi p} \mathfrak{W}_{i_{\xi+1} p} \prod_{j=1}^{\ell} \theta_3^{(j)2}$. Each element of the kind $\sigma_x^2 \sum_{p=1}^{n} \mathfrak{W}_{i_\xi p} \mathfrak{W}_{i_{\xi+1} p} \prod_{j=1}^{\ell} \theta_3^{(j)2}$ divides the sequence of products into independent blocks $\Pi_j^{(n,l)}$ as it was for the case in Appendix A.2. In this section we are also considering the covariance $\mathbf{\Sigma}^{(\ell)}$ and the contribution of each addend $E_\omega^{(n,k,n_w,n_\phi,p)}$ is considered in the following form

$$E_\omega^{(n,k,n_w,n_\phi,p)} = \int \prod_j E_{\Pi_i^{(n,n_\phi^{(j)})}} \mathcal{D}\mathbf{\Sigma}^{(\ell)} \tag{A.277}$$

and therefore in the following expression of $E_{\Pi_i^{(n,n_\phi^{(j)})}}$, $\mathbf{\Sigma}^{(\ell)}$ should be considered as a random variable.

Lemma A.3.2. *For the matrix*

$$\boldsymbol{M} = \frac{1}{m} \mathbb{E}_{\boldsymbol{X}} \left[\phi(\boldsymbol{W}^{(\ell)} \boldsymbol{Y}^{(\ell)}) \phi(\boldsymbol{W}^{(\ell)} \boldsymbol{Y}^{(\ell)})^\top \right]$$

$$- \frac{1}{m^2\sigma_x^2}\mathbb{E}_{\boldsymbol{x}}\left[\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})\boldsymbol{X}^\top\right]\mathbb{E}_{\boldsymbol{x}}\left[\phi(\boldsymbol{W}^{(\ell)}\boldsymbol{Y}^{(\ell)})\boldsymbol{X}^\top\right] \tag{A.278}$$

and $\boldsymbol{Y}^{(\ell)} = \phi\left(\boldsymbol{H}^{(\ell-1)}\right)$ *as defined in Theorem 3.3.5 without considering the elements in* $\boldsymbol{H}^{(\ell)}$ *independent, i.e.* $\boldsymbol{Y}^{(\ell)}_{:p} \sim \mathcal{N}(\boldsymbol{0}, q^{(\ell)}\boldsymbol{\Sigma}^{(\ell)})$*, when* $k > 1$ *each block* $\Pi_j^{(l)}$ *generates the following contribution*

$$E^{(n,\ell)}_{\Pi_i^{(l)}} = -\frac{\sigma_w^{2\ell}\sigma_x^2\left(\prod_{j=1}^{\ell}\theta_3^{(j)}\right)^2\tilde{\theta}_2^{(n)l}}{n^{1+l}}\left(\frac{1}{n}\sum_{k=1}^{n}{\boldsymbol{\Sigma}^{\ell}_{kk}}^{l}\right)\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.279}$$

with

$$\tilde{\theta}_2^{(n,\ell)} = \left(\int \frac{\sqrt{q^{(\ell)}}}{\sqrt{2\pi}}e^{-\frac{z^2}{2}}\phi'(\sqrt{q^{(\ell)}}\sqrt{\frac{tr(\boldsymbol{\Sigma}^{\ell})}{n}}z)dz\right)^2, \tag{A.280}$$

$$\theta_3 = \int_{z_1}\phi'\left(\sqrt{q^{(\ell)}}z_1\right)\mathcal{D}z_1 \tag{A.281}$$

and $\boldsymbol{\Sigma}^{(\ell)}$ *being the covariance matrix of the hidden layers* $\boldsymbol{Y}^{(\ell)}$*.*

Proof. Focusing on the integration of one independent block $\Pi_j^{(l)}$, where l identifies the numbers of factors $\phi(\sum_p \mathbf{W}^{(\ell)}_{i_\xi,p}\mathbf{Y}^{(\ell)}_{p,\mu_\xi})\phi(\sum_p \mathbf{W}^{(\ell)}_{i_{\xi+1},p}\mathbf{Y}^{(\ell)}_{p,\mu_\xi})$ between two of the $\sigma_x\left(\prod_{j=1}^{\ell}\theta_3^{(j)}\right)\mathfrak{W}^{(\ell)}_{i_\xi p}$ kind, i.e.

$$E^{(n,\ell)}_{\Pi_i^{(l)}} = -\int_{\mathbf{W},\mathbf{Y}}\left\{\sigma_x\left(\prod_{j=1}^{\ell}\theta_3^{(j)}\right)\mathfrak{W}^{(\ell)}_{i_1 p}\prod_{\xi=1}^{l}\left(\phi(\sum_p \mathbf{W}^{(\ell)}_{i_\xi,p}\mathbf{Y}^{(\ell)}_{p,\mu_\xi})\phi(\sum_p \mathbf{W}^{(\ell)}_{i_{\xi+1},p}\mathbf{Y}^{(\ell)}_{p,\mu_\xi})\right)\right.$$
$$\left.\cdot\sigma_x\left(\prod_{j=1}^{\ell}\theta_3^{(j)}\right)\mathfrak{W}^{(\ell)}_{i_{l+1}q}\right\}\mathcal{D}\mathbf{W}\mathcal{D}\mathbf{Y}. \tag{A.282}$$

By defining as $\boldsymbol{\Sigma}^{(\ell)}$ the covariance of the post-activation layer, we are now able to consider the following equality

$$\sum_p \mathbf{W}^{(\ell)}_{i_\xi,p}\mathbf{Y}^{(\ell)}_{p,\mu_\xi} = \sum_p \tilde{\mathbf{W}}^{(\ell)}_{i_\xi,p}{\boldsymbol{\Sigma}^{(\ell)}_{pp}}^{1/2}\tilde{\mathbf{Y}}_{p,\mu_\xi} \tag{A.283}$$

where $\tilde{\mathbf{W}}^{(\ell)} \sim \mathbf{W}^{(\ell)}$ and $\tilde{\mathbf{Y}}_{ij} \sim \mathcal{N}\left(0, \frac{q^{(\ell)}}{\sigma_w^2}\right)$ and this allows to run analysis similar to Appendix A.2.

The computation of this expectation follows the structure of the proof in [72]. A dummy variable z is introduced with a delta Dirac function within each ϕ element, and then a Fourier representation for all the arguments of the ϕ functions is introduced. For each factor Π_j the set $\mathcal{Z}_{\Pi_j} \subset \mathcal{Z}$ is considered containing only the combinations (i_ξ, μ_ν) that are included in the $\phi(\sum_l \tilde{\mathbf{W}}_{i_\xi,l} {\boldsymbol{\Sigma}^{(\ell)}_{ll}}^{1/2} \tilde{\mathbf{Y}}_{l,\mu_\nu})$ arguments. Thus auxiliary integrals over z are considered by adding delta functions enforcing $\mathbf{Z} = \mathbf{W}^{(\ell)}\mathbf{Y}^{(\ell)} = \tilde{\mathbf{W}}^{(\ell)} {\boldsymbol{\Sigma}^{(\ell)}}^{1/2} \tilde{\mathbf{Y}}$ with

$$\mathbf{Z}_{i\mu} = \begin{cases} z_{i\mu} & \text{if } \ (i,\mu) \in \mathcal{Z}_{\Pi_j} \\ 0 & \text{otherwise.} \end{cases} \tag{A.284}$$

and consequently

$$\begin{aligned} E^{(n,\ell)}_{\Pi_i^{(l)}} = &- \int \left\{ \sigma_x \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right) \mathfrak{W}_{i_1 p} \prod_{\xi=1}^{l} \left(\phi(\sum_p \tilde{\mathbf{W}}^{(\ell)}_{i_\xi,p} {\boldsymbol{\Sigma}^{(\ell)}_{pp}}^{1/2} \tilde{\mathbf{Y}}_{p,\mu_\xi}) \phi(\sum_p \tilde{\mathbf{W}}^{(\ell)}_{i_{\xi+1},p} {\boldsymbol{\Sigma}^{(\ell)}_{pp}}^{1/2} \tilde{\mathbf{Y}}_{p,\mu_\xi}) \right) \right. \\ &\left. \cdot \sigma_x \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right) \mathfrak{W}_{i_{l+1} q} \right\} \mathcal{D}\mathbf{W} \mathcal{D}\tilde{\mathbf{Y}} \end{aligned} \tag{A.285}$$

$$\begin{aligned} = &- \int \prod_{z_{\alpha\beta} \in \mathcal{Z}} \delta(z_{\alpha\beta} - \sum_k \tilde{\mathbf{W}}^{(\ell)}_{\alpha k} {\boldsymbol{\Sigma}^{(\ell)}_{kk}}^{1/2} \tilde{\mathbf{Y}}_{k\beta}) \cdot \\ &\cdot \left\{ \sigma_x^2 \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right)^2 \mathfrak{W}_{i_1 p} \prod_{\xi=1}^{l} \left(\phi(z_{i_\xi \mu_\xi}) \phi(z_{i_\xi \mu_\xi}) \right) \mathfrak{W}_{i_{l+1} q} \right\} \mathcal{D}\mathbf{W} \mathcal{D}\tilde{\mathbf{Y}} \mathcal{D}z \end{aligned} \tag{A.286}$$

$$\begin{aligned} = &- \int e^{-itr\boldsymbol{\Lambda}^\top (\tilde{\mathbf{W}}^{(\ell)} {\boldsymbol{\Sigma}^{(\ell)}}^{1/2} \tilde{\mathbf{Y}} - \mathbf{Z})} \sigma_x^2 \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right)^2 \cdot \\ &\cdot \mathfrak{W}_{i_1 p} \prod_{\xi=1}^{l} \left(\phi(z_{i_\xi \mu_\xi}) \phi(z_{i_{\xi+1} \mu_\xi}) \right) \mathfrak{W}_{i_{l+1} q} \mathcal{D}\mathbf{W} \mathcal{D}\tilde{\mathbf{Y}} \mathcal{D}\lambda \mathcal{D}z \end{aligned} \tag{A.287}$$

where

$$\mathcal{D}z = \prod_{(\alpha,\beta) \in \mathcal{Z}_{\Pi_i}} dz_{\alpha\beta} \tag{A.288}$$

and in the second equality the following property is used

$$\delta(x) = \frac{1}{2\pi} \int e^{i\lambda x} d\lambda \tag{A.289}$$

and therefore the matrix $\mathbf{\Lambda} \in \mathbb{R}^{n \times m}$ whose entries are

$$\mathbf{\Lambda}_{i\mu} = \begin{cases} \lambda_{i\mu} & \text{if } (i,\mu) \in \mathcal{Z}_{\Pi_i} \\ 0 & \text{otherwise.} \end{cases} \tag{A.290}$$

is considered and therefore

$$\mathcal{D}\lambda = \prod_{(\alpha,\beta)\in\mathcal{Z}_{\Pi_i}} \frac{d\lambda_{\alpha\beta}}{2\pi}. \tag{A.291}$$

To start with, the integration over $\tilde{\mathbf{Y}}$ is considered

$$\int e^{-itr(\mathbf{\Lambda}^\top \tilde{\mathbf{W}}^{(\ell)} \mathbf{\Sigma}^{(\ell)\,1/2} \tilde{\mathbf{Y}})} \mathcal{D}\tilde{\mathbf{Y}} = \prod_{b,c=1}^{m,n} \int \frac{d\tilde{\mathbf{Y}}_{cb}}{\sqrt{2\pi q^{(\ell)}}} exp\left[-\frac{\sigma_w^2}{2q^\ell} \tilde{\mathbf{Y}}_{cb}^2 - i \sum_{a=1}^{n} \lambda_{ab} \tilde{\mathbf{W}}_{ac}^{(\ell)} \mathbf{\Sigma}_{cc}^{(\ell)\,1/2} \tilde{\mathbf{Y}}_{cb} \right]$$

$$= \prod_{b,c=1}^{m,n} \int \frac{d\tilde{\mathbf{Y}}_{cb}}{\sqrt{2\pi q^{(\ell)}}} exp\left[-\frac{\sigma_w^2}{2q^{(\ell)}} \left(\tilde{\mathbf{Y}}_{cb} + i \frac{q^{(\ell)}}{\sigma_w^2} (\sum_{a=1}^{n} \lambda_{ab} \tilde{\mathbf{W}}_{ac}^{(\ell)} \mathbf{\Sigma}_{cc}^{(\ell)\,1/2}) \right)^2 - \frac{q^{(\ell)}}{2\sigma_w^2} (\sum_{a=1}^{n} \lambda_{ab} \tilde{\mathbf{W}}_{ac}^{(\ell)})^2 \right] \tag{A.292}$$

$$= \prod_{b,c=1}^{m,n} exp\left[-\frac{q^{(\ell)}}{2\sigma_w^2} (\sum_{a=1}^{n} \lambda_{ab} \tilde{\mathbf{W}}_{ac}^{(\ell)} \mathbf{\Sigma}_{cc}^{(\ell)\,1/2})^2 \right] = exp\left[-\frac{q^{(\ell)}}{2\sigma_w^2} \sum_{b,c=1}^{m,n} (\sum_{a=1}^{n} \lambda_{ab} \tilde{\mathbf{W}}_{ac}^{(\ell)} \mathbf{\Sigma}_{cc}^{(\ell)\,1/2})^2 \right] \tag{A.293}$$

$$= exp\left[-\frac{q^{(\ell)}}{2\sigma_w^2} \| \mathbf{\Lambda}^\top \tilde{\mathbf{W}}^{(\ell)} \mathbf{\Sigma}^{(\ell)\,1/2} \|_F^2 \right] = e^{-\frac{q^{(\ell)}}{2\sigma_w^2} tr(\mathbf{\Sigma}^{(\ell)\top\,1/2} \tilde{\mathbf{W}}^{(\ell)\top} \mathbf{\Lambda}\mathbf{\Lambda}^\top \tilde{\mathbf{W}}^{(\ell)} \mathbf{\Sigma}^{(\ell)\,1/2})}. \tag{A.294}$$

and this allows to integrate over $\tilde{\mathbf{W}}^{(\ell)}$. Note that $\tilde{\mathbf{W}}^{(\ell)} = \mathbf{W}^{(\ell)} \mathbf{O}^\top$ and the elements in $\mathfrak{W}^{(\ell)}$ are dependent on $\mathbf{W}^{(\ell)}$, therefore it is also necessary to integrate over the possible orthonormal transform $\mathbf{O}$. The case where $p \neq q$ is now considered

$$\int_{\tilde{\mathbf{W}},\mathbf{O}} \sigma_x^2 \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right)^2 e^{-\frac{q^{(\ell)}}{2\sigma_w^2} tr(\mathbf{\Sigma}^{(\ell)} \tilde{\mathbf{W}}^{(\ell)\top} \mathbf{\Lambda}\mathbf{\Lambda}^\top \mathbf{W}^{(\ell)})} \mathfrak{W}_{i_1 p}^{(\ell)} \mathfrak{W}_{i_{l+1} q}^{(\ell)} \mathcal{D}\mathbf{W} \mathcal{D}\mathbf{O}$$

$$= \sigma_x^2 \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right)^2 \int \left(\prod_{i,j=1}^{n} \frac{d\tilde{\mathbf{W}}_{ij}^{(\ell)}}{\sqrt{2\pi\sigma_w^2/n}} \right) e^{-\frac{ntr(\tilde{\mathbf{W}}^{(\ell)\top} \tilde{\mathbf{W}}^{(\ell)})}{2\sigma_w^2}} e^{-\frac{q^{(\ell)}}{2\sigma_w^2} tr(\mathbf{\Sigma}^{(\ell)} \tilde{\mathbf{W}}^{(\ell)\top} \mathbf{\Lambda}\mathbf{\Lambda}^\top \tilde{\mathbf{W}}^{(\ell)})} \mathfrak{W}_{i_1 p}^{(\ell)} \mathfrak{W}_{i_{l+1} q}^{(\ell)} \tag{A.295}$$

$$= \sigma_x^2 \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right)^2 \int \left(\prod_{j=1}^{n} \frac{d^n \tilde{\mathbf{w}}_j}{(2\pi\sigma_w^2/n)^{n/2}} \right) e^{-\sum_{j=1}^{n} (\frac{n}{2\sigma_w^2} \tilde{\mathbf{w}}_j^\top \tilde{\mathbf{w}}_j + \frac{q^{(\ell)}}{2\sigma_w^2} \mathbf{\Sigma}_{jj}^{(\ell)} \tilde{\mathbf{w}}_j^\top \mathbf{\Lambda}\mathbf{\Lambda}^\top \tilde{\mathbf{w}}_j)}.$$

$$\cdot \left(\sum_{\alpha,\beta=1}^{n} \mathbf{w}_{\alpha}^{(i_1)} \mathbf{w}_{\beta}^{(i_{l+1})} \mathfrak{W}_{\alpha p}^{(\ell-1)} \mathfrak{W}_{\beta q}^{(\ell-1)} \right) \tag{A.296}$$

$$= \sigma_x^2 \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right)^2 \left(\sum_{\alpha,\beta=1}^{n} \int \left(\prod_{j=1}^{n} \frac{d^n \tilde{\mathbf{w}}_j}{(2\pi\sigma_w^2/n)^{n/2}} \right) e^{-\sum_{j=1}^{n} (\frac{n}{2\sigma_w^2} \tilde{w}_j^\top \tilde{\mathbf{w}}_j + \frac{q^{(\ell)}}{2\sigma_w^2} \mathbf{\Sigma}_{jj}^{(\ell)} \tilde{\mathbf{w}}_j^\top \mathbf{\Lambda}\mathbf{\Lambda}^\top \tilde{\mathbf{w}}_j)} \cdot \right.$$

$$\left. \cdot \, \mathbf{w}_{\alpha}^{(i_1)} \mathbf{w}_{\beta}^{(i_{l+1})} \underbrace{\int \mathfrak{W}_{\alpha p}^{(\ell-1)} \mathfrak{W}_{\beta q}^{(\ell-1)}}_{=0} \right) \tag{A.297}$$

$$= 0. \tag{A.298}$$

While now the case where $q = p$ will be considered

$$\int_{\tilde{\mathbf{W}},\mathbf{O}} \sigma_x^2 \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right)^2 e^{-\frac{\sigma_x^2}{2} tr(\mathbf{\Sigma}^{(\ell)} \tilde{\mathbf{W}}^{(\ell)\top} \mathbf{\Lambda}\mathbf{\Lambda}^\top \mathbf{W}^{(\ell)})} \mathfrak{W}_{i_1 p}^{(\ell)} \mathfrak{W}_{i_{l+1} p}^{(\ell)} \mathcal{D}\mathbf{W} \mathcal{D}\mathbf{O}$$

$$= \sigma_x^2 \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right)^2 \int \left(\prod_{i,j=1}^{n} \frac{d\tilde{\mathbf{W}}_{ij}^{(\ell)}}{\sqrt{2\pi\sigma_w^2/n}} \right) e^{-\frac{ntr(\tilde{\mathbf{W}}^{(\ell)\top} \tilde{\mathbf{W}}^{(\ell)})}{2\sigma_w^2}} e^{-\frac{q^{(\ell)}}{2\sigma_w^2} tr(\mathbf{\Sigma}^{(\ell)} \tilde{\mathbf{W}}^{(\ell)\top} \mathbf{\Lambda}\mathbf{\Lambda}^\top \tilde{\mathbf{W}}^{(\ell)})}$$

$$\cdot \, \mathfrak{W}_{i_1 p}^{(\ell)} \mathfrak{W}_{i_{l+1} p}^{(\ell)} \mathcal{D}\mathbf{O} \tag{A.299}$$

$$= \sigma_x^2 \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right)^2 \int \left(\prod_{j=1}^{n} \frac{d^n \tilde{\mathbf{w}}_j}{(2\pi\sigma_w^2/n)^{n/2}} \right) e^{-\sum_{j=1}^{n} (\frac{n}{2\sigma_w^2} \tilde{\mathbf{w}}_j^\top \tilde{\mathbf{w}}_j + \frac{q^{(\ell)}}{2\sigma_w^2} \mathbf{\Sigma}_{jj} \tilde{\mathbf{w}}_j^\top \mathbf{\Lambda}\mathbf{\Lambda}^\top \tilde{\mathbf{w}}_j)}$$

$$\cdot \left(\sum_{\alpha,\beta=1}^{n} \mathbf{w}_{\alpha}^{(i_1)} \mathfrak{W}_{\alpha p}^{(\ell-1)} \mathbf{w}_{\beta}^{(i_{l+1})} \mathfrak{W}_{\beta p}^{(\ell-1)} \right) \mathcal{D}\mathbf{O} \tag{A.300}$$

$$= \sigma_x^2 \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right)^2 \sum_{\alpha=1}^{n} \left(\frac{\sigma_w^{2(\ell-1)}}{n} \int \left(\prod_{j=1}^{n} \frac{d^n \tilde{\mathbf{w}}_j}{(2\pi\sigma_w^2/n)^{n/2}} \right) e^{-\sum_{j=1}^{n} (\frac{n}{2\sigma_w^2} \tilde{\mathbf{w}}_j^\top \tilde{\mathbf{w}}_j + \frac{q^{(\ell)}}{2\sigma_w^2} \mathbf{\Sigma}_{jj} \tilde{\mathbf{w}}_j^\top \mathbf{\Lambda}\mathbf{\Lambda}^\top \tilde{\mathbf{w}}_j)} \right.$$

$$\left. \cdot \, \mathbf{w}_{\alpha}^{(i_1)} \mathbf{w}_{\alpha}^{(i_{l+1})} \right) \mathcal{D}\mathbf{O} \tag{A.301}$$

$$= \sigma_x^2 \sigma_w^{2(\ell-1)} \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right)^2 \int \left(\prod_{j=1}^{n} \frac{d^n \tilde{\mathbf{w}}_j}{(2\pi\sigma_w^2/n)^{n/2}} \right) e^{-\sum_{j=1}^{n} (\frac{n}{2\sigma_w^2} \tilde{\mathbf{w}}_j^\top \tilde{\mathbf{w}}_j + \frac{q^{(\ell)}}{2\sigma_w^2} \mathbf{\Sigma}_{jj} \tilde{\mathbf{w}}_j^\top \mathbf{\Lambda}\mathbf{\Lambda}^\top \tilde{\mathbf{w}}_j)} \cdot$$

$$\cdot \, (\sum_{k=1}^{n} \mathbf{O}_{\alpha k} \tilde{\mathbf{w}}_k^{(i_1)}) (\sum_{k=1}^{n} \mathbf{O}_{\alpha k} \tilde{\mathbf{w}}_k^{(i_{l+1})}) \mathcal{D}\mathbf{O} \tag{A.302}$$

$$= \sum_{k=1}^{n} \sigma_x^2 \sigma_w^{2(\ell-1)} \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right)^2 \cdot$$

$$\cdot \int \left(\prod_{j=1}^{n} \frac{d^n \tilde{\mathbf{w}}_j}{(2\pi\sigma_w^2/n)^{n/2}} \right) e^{-\sum_{j=1}^{n} (\frac{n}{2\sigma_w^2} \tilde{\mathbf{w}}_j^\top \tilde{\mathbf{w}}_j + \frac{q^{(\ell)}}{2\sigma_w^2} \mathbf{\Sigma}_{jj} \tilde{\mathbf{w}}_j^\top \mathbf{\Lambda}\mathbf{\Lambda}^\top \tilde{\mathbf{w}}_j)} \tilde{\mathbf{W}}_k^{(i_1)} \tilde{\mathbf{W}}_k^{(i_{l+1})}$$

$$\cdot \underbrace{\int_{\mathbf{O}} \mathbf{O}_{\alpha k}^2 \mathcal{D}\mathbf{O}}_{=1/n} \tag{A.303}$$

$$= \sum_{k=1}^{n} \frac{\sigma_x^2 \sigma_w^{2(\ell-1)}}{n} \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right)^2 \int \left(\prod_{\substack{j=1 \\ j\neq k}}^{n} \frac{d^n \tilde{\mathbf{w}}_j}{(2\pi\sigma_w^2/n)^{n/2}} \right) e^{-\sum_{\substack{j=1 \\ j\neq k}}^{n} (\frac{n}{2\sigma_w^2} \tilde{\mathbf{w}}_j^\top \tilde{\mathbf{w}}_j + \frac{q^{(\ell)} \mathbf{\Sigma}_{jj}}{2\sigma_w^2} \tilde{\mathbf{w}}_j^\top \mathbf{\Lambda}\mathbf{\Lambda}^\top \tilde{\mathbf{w}}_j)}$$

$$\cdot \int \frac{d^n \tilde{\mathbf{w}}_k}{(2\pi\sigma_w^2/n)^{n/2}} \tilde{\mathbf{w}}_k^{(i_1)} \tilde{\mathbf{w}}_k^{(i_{l+1})} e^{-\frac{n}{2\sigma_w^2} \tilde{\mathbf{w}}_k^\top \tilde{\mathbf{w}}_k - \frac{q^{(\ell)} \mathbf{\Sigma}_{kk}}{2\sigma_w^2} \tilde{\mathbf{w}}_k^\top \mathbf{\Lambda}\mathbf{\Lambda}^\top \tilde{\mathbf{w}}_k} \tag{A.304}$$

$$= \sum_{k=1}^{n} \frac{\sigma_x^2 \sigma_w^{2(\ell-1)}}{n} \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right)^2 \cdot$$

$$\cdot \prod_{\substack{j=1 \\ j\neq k}}^{n} \underbrace{\left(\int \frac{d^n \tilde{\mathbf{w}}_j}{(2\pi\sigma_w^2/n)^{n/2}} \frac{det^{-1}(\mathbf{I} + \frac{q^{(\ell)} \mathbf{\Sigma}_{jj}}{n} \mathbf{\Lambda}\mathbf{\Lambda}^\top)^{1/2}}{det^{-1}(\mathbf{I} + \frac{q^{(\ell)} \mathbf{\Sigma}_{jj}}{n} \mathbf{\Lambda}\mathbf{\Lambda}^\top)^{1/2}} e^{-\frac{n}{2\sigma_w^2} \tilde{\mathbf{w}}_j^\top \tilde{\mathbf{w}}_j - \frac{q^{(\ell)} \mathbf{\Sigma}_{jj}}{2\sigma_w^2} \tilde{\mathbf{w}}_j^\top \mathbf{\Lambda}\mathbf{\Lambda}^\top \tilde{\mathbf{w}}_j)} \right)}_{= \frac{1}{det(\mathbf{I} + \frac{q^{(\ell)} \mathbf{\Sigma}_{jj}}{n} \mathbf{\Lambda}\mathbf{\Lambda}^\top)^{1/2}}}$$

$$\cdot \int \frac{d^n \tilde{\mathbf{w}}_k}{(2\pi\sigma_w^2/n)^{n/2}} \frac{det^{-1}(\mathbf{I} + \frac{q^{(\ell)} \mathbf{\Sigma}_{kk}}{n} \mathbf{\Lambda}\mathbf{\Lambda}^\top)^{1/2}}{det^{-1}(\mathbf{I} + \frac{q^{(\ell)} \mathbf{\Sigma}_{kk}}{n} \mathbf{\Lambda}\mathbf{\Lambda}^\top)^{1/2}} \tilde{\mathbf{w}}_k^{(i_0)} \tilde{\mathbf{w}}_k^{(i_{l+1})} e^{-\frac{1}{2} \tilde{\mathbf{w}}_k^\top \left(\frac{n}{\sigma_w^2} I + \frac{q^{(\ell)} \mathbf{\Sigma}_{kk}}{\sigma_w^2} \mathbf{\Lambda}\mathbf{\Lambda}^\top \right) \tilde{\mathbf{w}}_k} \tag{A.305}$$

$$= \sum_{k=1}^{n} \frac{\sigma_x^2 \sigma_w^{2(\ell-1)} \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right)^2}{n \prod_{j=1}^{n} det(\mathbf{I} + \frac{q^{(\ell)} \mathbf{\Sigma}_{jj}}{n} \mathbf{\Lambda}\mathbf{\Lambda}^\top)^{1/2}} Cov \left[\tilde{\mathbf{w}}_k^{(i_0)} \tilde{\mathbf{w}}_k^{(i_{l+1})} \right]. \tag{A.306}$$

The covariance matrix of the vector $\tilde{\mathbf{w}}_k$ is

$$\left(\frac{n}{\sigma_w^2} \mathbf{I} + \frac{q^{(\ell)} \mathbf{\Sigma}_{kk}}{\sigma_w^2} \mathbf{\Lambda}\mathbf{\Lambda}^\top \right)^{-1} = \frac{\sigma_w^2}{n} \sum_{k=0}^{\infty} (-\frac{q^{(\ell)} \mathbf{\Sigma}_{kk}}{n} \mathbf{\Lambda}\mathbf{\Lambda}^\top)^k \tag{A.307}$$

therefore

$$Cov \left[\tilde{\mathbf{w}}_k^{(i_0)} \tilde{\mathbf{w}}_k^{(i_{l+1})} \right] = \left[\left(\frac{n}{\sigma_w^2} I + \frac{q^{(\ell)} \mathbf{\Sigma}_{kk}}{\sigma_w^2} \mathbf{\Lambda}\mathbf{\Lambda}^\top \right)^{-1} \right]_{i_0 i_{l+1}} = \left[\frac{\sigma_w^2}{n} \sum_{\nu=0}^{\infty} (-\frac{q^{(\ell)} \mathbf{\Sigma}_{kk}}{n} \mathbf{\Lambda}\mathbf{\Lambda}^\top)^\nu \right]_{i_0 i_{l+1}}$$

$$\tag{A.308}$$

$$= \frac{\sigma_w^2}{n} \sum_{\nu=l}^{\infty} \left[(-\frac{q^{(\ell)} \mathbf{\Sigma}_{kk}}{n} \mathbf{\Lambda}\mathbf{\Lambda}^\top)^\nu \right]_{i_0 i_{l+1}} \tag{A.309}$$

$$= (-1)^l \frac{\sigma_w^2}{n} \prod_{j=1}^{l} \frac{q^{(\ell)} \mathbf{\Sigma}_{kk}}{n} \lambda_{i_j \mu_j} \lambda_{i_{j+1} \mu_j} + \sum_{\nu=1}^{\infty} (-1)^{l+\nu} \frac{\sigma_w^2}{n} \sum_{\substack{\overrightarrow{l} \, s.t. \\ \| \overrightarrow{l} \|_1 = \nu + l \\ \tilde{l}_j \in \mathbb{N}_0}} \prod_{j=1}^{l} \left(\frac{q^{(\ell)} \mathbf{\Sigma}_{kk}}{n} \lambda_{i_j \mu_j} \lambda_{i_{j+1} \mu_j} \right)^{\tilde{l}_j}. \tag{A.310}$$

Exactly as shown in Lemma 3.3.3, the first addend in (A.310) determines the contribution $E^{(n,\ell)}_{\Pi_i^{(l)}}$ up to a relative error of $\mathcal{O}(1/n)$. This implies that by considering $F(z) = \prod_{\xi=1}^{l} \left(\phi(z_{i_\xi \mu_\xi}) \phi(z_{i_{\xi+1} \mu_\xi}) \right)$

$$E^{(n,\ell)}_{\Pi_i^{(l)}} = - \int e^{-itr\mathbf{\Lambda}^\top (\tilde{\mathbf{W}}^{(\ell)} \mathbf{\Sigma}^{(\ell) 1/2} \tilde{\mathbf{Y}} - \mathbf{Z})} \sigma_x^2 \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right)^2 \cdot$$

$$\cdot \mathfrak{W}_{i_1 p} \prod_{\xi=1}^{l} \left(\phi(z_{i_\xi \mu_\xi}) \phi(z_{i_{\xi+1} \mu_\xi}) \right) \mathfrak{W}_{i_{l+1} q} \mathcal{D}\mathbf{W} \mathcal{D}X \mathcal{D}\lambda \mathcal{D}z \tag{A.311}$$

$$= - \sum_{k=1}^{n} \int \frac{\sigma_x^2 \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right)^2 \sigma_w^{2(\ell-1)}}{n} (-1)^l \frac{\sigma_w^2}{n} \left(\prod_{j=1}^{l} \frac{q^{(\ell)} \mathbf{\Sigma}^{(\ell)}_{kk}}{n} \lambda_{i_j \mu_j} \lambda_{i_{j+1} \mu_j} \right)$$

$$\frac{1}{\prod_{j=1}^{n} det(I + \frac{q^{(\ell)} \mathbf{\Sigma}^{(\ell)}_{jj}}{n} \mathbf{\Lambda}\mathbf{\Lambda}^\top)^{1/2}} e^{tr\mathbf{\Lambda}^\top \mathbf{Z}} F(z) \mathcal{D}\lambda \mathcal{D}z \left(1 + \mathcal{O}\left(\frac{1}{n} \right) \right) \tag{A.312}$$

$$= - \sum_{k=1}^{n} \int \frac{\sigma_x^2 \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right)^2}{n} (-1)^l \frac{\sigma_w^{2\ell}}{n} \left(\prod_{j=1}^{l} \frac{q^{(\ell)} \mathbf{\Sigma}^{(\ell)}_{kk}}{n} \lambda_{i_j \mu_j} \lambda_{i_{j+1} \mu_j} \right)$$

$$e^{- \sum_{j=1}^{n} \frac{1}{2} \log det(I + \frac{q^{(\ell)} \mathbf{\Sigma}^{(\ell)}_{jj}}{n} \mathbf{\Lambda}\mathbf{\Lambda}^\top)} e^{tr\mathbf{\Lambda}^\top \mathbf{Z}} F(z) \mathcal{D}\lambda \mathcal{D}z \left(1 + \mathcal{O}\left(\frac{1}{n} \right) \right) \tag{A.313}$$

$$= -(-1)^l \sum_{k=1}^{n} \frac{\sigma_x^2 \sigma_w^{2\ell} \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right)^2}{n^2} \int \left(\prod_{(\alpha,\beta) \in \mathcal{Z}_{\Pi_j}} \frac{\sqrt{q^{(\ell)} \mathbf{\Sigma}^{(\ell)}_{kk}}}{\sqrt{n}} \lambda_{\alpha\beta} \right)$$

$$e^{- \frac{q^{(\ell)} tr \mathbf{\Sigma}^{(\ell)}}{2n} tr(\mathbf{\Lambda}\mathbf{\Lambda}^\top) - \frac{1}{2} \sum_{\xi \geq 2} \frac{(-1)^{\xi+1}}{\xi} tr \mathbf{\Sigma}^{(\ell)\xi} tr \left(\frac{q^{(\ell)}}{n} \mathbf{\Lambda}\mathbf{\Lambda}^\top \right)^\xi - itr\mathbf{\Lambda}^\top \mathbf{Z}} F(z) \mathcal{D}\lambda \mathcal{D}z$$

$$\left(1 + \mathcal{O}\left(\frac{1}{n} \right) \right) \tag{A.314}$$

where in the last equation the Taylor expansion $logdet|\mathbf{I}+\mathbf{X}| = \sum_{\xi=1} \frac{(-1)^{\xi+1}}{\xi} tr(\mathbf{X}^{\xi})$ was used. Similar to Lemma 3.3.3, the first order expansion of the log-determinant. $\xi = 1$, is the leading order contribution with a relative error of $\mathcal{O}(1/n)$. By introducing the following change of variable,

$$\tilde{\lambda}_{\alpha\beta} = \sqrt{q^{(\ell)}}\lambda_{\alpha\beta} \tag{A.315}$$

we then retrieve the following expression

$$\begin{aligned}
E^{(n,\ell)}_{\Pi_i^{(\ell)}} &= -(-1)^l \sum_{k=1}^{n} \frac{\sigma_w^2\sigma_x^2 \left(\prod_{j=1}^{\ell}\theta_3^{(j)}\right)^2 \mathbf{\Sigma}^{(\ell)l}_{kk}}{n^2} \cdot \\
&\quad \cdot \int \left(\prod_{(\alpha,\beta)\in\mathcal{Z}_{\Pi_j}} \frac{\sqrt{q^{(\ell)}}}{\sqrt{n}}\lambda_{\alpha\beta}\right) e^{-\frac{q^{(\ell)}tr\mathbf{\Sigma}}{2n} tr(\mathbf{\Lambda}\mathbf{\Lambda}^\top) - itr\mathbf{\Lambda}^\top\mathbf{Z}} F(z)\mathcal{D}\lambda\mathcal{D}z \\
&\quad \left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \qquad (A.316) \\
&= -(-1)^l \sum_{k=1}^{n} \frac{\sigma_x^2\sigma_w^{2\ell}\left(\prod_{j=1}^{\ell}\theta_3^{(j)}\right)^2 \mathbf{\Sigma}^{(\ell)l}_{kk}}{n^2} \int \left(\prod_{(\alpha,\beta)\in\mathcal{Z}_{\Pi_j}} \frac{\tilde{\lambda}_{\alpha\beta}}{\sqrt{n}}\right) \\
&\qquad e^{-\sum_{(\alpha,\beta)\in\mathcal{Z}_{\Pi_j}}\left(\frac{tr(\mathbf{\Sigma})\tilde{\lambda}^2_{\alpha\beta}}{2n} + i\frac{\tilde{\lambda}_{\alpha\beta}z_{\alpha\beta}}{\sqrt{q^{(\ell)}}}\right)} F(z) \left(\prod_{(\alpha,\beta)\in\mathcal{Z}_{\Pi_j}} \frac{d\tilde{\lambda}_{\alpha\beta}}{2\pi\sqrt{q^{(\ell)}}}\right) \\
&\qquad \left(\prod_{(\alpha,\beta)\in\mathcal{Z}_{\Pi_j}} dz_{\alpha\beta}\right)\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \qquad (A.317) \\
&= -(-1)^l \sum_{k=1}^{n} \frac{\sigma_x^2\sigma_w^{2\ell}\left(\prod_{j=1}^{\ell}\theta_3^{(j)}\right)^2 \mathbf{\Sigma}^{(\ell)l}_{kk}}{n^2} \cdot \\
&\quad \cdot \prod_{(\alpha\beta)\in\mathcal{Z}_{\Pi_j}} \left(\int \frac{\tilde{\lambda}_{\alpha\beta}}{\sqrt{n}} e^{-\frac{tr(\mathbf{\Sigma})\tilde{\lambda}^2_{\alpha\beta}}{2n} - i\frac{\tilde{\lambda}_{\alpha\beta}z_{\alpha\beta}}{\sqrt{q^{(\ell)}}}} \phi(z_{\alpha\beta}) \frac{d\tilde{\lambda}_{\alpha\beta}}{2\pi\sqrt{q^{(\ell)}}} dz_{\alpha\beta}\right) \\
&\qquad \left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \qquad (A.318) \\
&= -(-1)^l \sum_{k=1}^{n} \frac{\sigma_x^2\sigma_w^{2\ell}\left(\prod_{j=1}^{\ell}\theta_3^{(j)}\right)^2 \mathbf{\Sigma}^{(\ell)l}_{kk}}{n^2}
\end{aligned}$$

$$\left(\int \frac{\tilde{\lambda}_{\alpha\beta}}{\sqrt{n}} e^{-\frac{1}{2}\left(\sqrt{\frac{tr(\mathbf{\Sigma})}{n}}\tilde{\lambda}_{\alpha\beta}+\frac{iz_{\alpha\beta}\sqrt{n}}{\sqrt{tr(\mathbf{\Sigma})}\sqrt{q^{(\ell)}}}\right)^2-\frac{z_{\alpha\beta}^2 n}{tr(\mathbf{\Sigma})2q^{(\ell)}}}\phi(z_{\alpha\beta})\frac{d\tilde{\lambda}_{\alpha\beta}}{2\pi\sqrt{q^{(\ell)}}}dz_{\alpha\beta}\right)^{2l}$$
$$\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.319}$$

$$= -(-1)^l \sum_{k=1}^{n} \frac{\sigma_x^2\sigma_w^{2\ell}\left(\prod_{j=1}^{\ell}\theta_3^{(j)}\right)^2 \mathbf{\Sigma}_{kk}^{(\ell)l}}{n^2}$$
$$\left(\int \frac{\phi(z)}{\sqrt{2\pi}\sqrt{n}\sqrt{q^{(\ell)}}} e^{-\frac{z^2 n}{2q^{(\ell)}tr\mathbf{\Sigma}}}\left(\int \frac{\tilde{\lambda}}{\sqrt{2\pi}} e^{-\frac{1}{2}\left(\sqrt{\frac{tr(\mathbf{\Sigma})}{n}}\tilde{\lambda}+\frac{iz\sqrt{n}}{\sqrt{tr(\mathbf{\Sigma})}\sqrt{q^{(\ell)}}}\right)^2} d\tilde{\lambda}\right) dz\right)^{2l}$$
$$\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.320}$$

$$= -(-1)^l \sum_{k=1}^{n} \frac{\sigma_x^2\sigma_w^{2\ell}\left(\prod_{j=1}^{\ell}\theta_3^{(j)}\right)^2 \mathbf{\Sigma}_{kk}^{(\ell)l}}{n^2}$$
$$\left(\int \frac{\phi(z)}{\sqrt{2\pi}\sqrt{n}\sqrt{q^{(\ell)}}} e^{-\frac{z^2 n}{2tr(\mathbf{\Sigma})q^{(\ell)}}}\left(-\frac{iz\sqrt{n}}{\sqrt{tr(\mathbf{\Sigma})}\sqrt{q^{(\ell)}}}\right) dz\right)^{2l}$$
$$\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.321}$$

$$= -(-1)^l \sum_{k=1}^{n} \frac{\sigma_x^2\sigma_w^{2\ell}\left(\prod_{j=1}^{\ell}\theta_3^{(j)}\right)^2 \mathbf{\Sigma}_{kk}^{(\ell)l}}{n^2}\left(\frac{-1}{n}\right)^l$$
$$\left(\int_{\mathbf{Z}} \frac{\phi(z)}{\sqrt{2\pi}} e^{-\frac{z^2 n}{2q^{(\ell)}tr(\mathbf{\Sigma})}} \frac{z\sqrt{n}}{\sqrt{tr(\mathbf{\Sigma})}q^{(\ell)}} dz\right)^{2l}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right). \tag{A.322}$$

Now the following change of variable is considered

$$\tilde{z} = \frac{z\sqrt{n}}{\sqrt{tr(\mathbf{\Sigma}^{(\ell)})}\sqrt{q^{(\ell)}}} \tag{A.323}$$

and then

$$E_{\Pi_j^{(\ell)}}^{(n,\ell)} = -\frac{\sigma_x^2\sigma_w^{2\ell}\left(\prod_{j=1}^{\ell}\theta_3^{(j)}\right)^2}{n^{1+l}}\left(\frac{1}{n}\sum_{k=1}^{n}\mathbf{\Sigma}_{kk}^{(\ell)l}\right)\left(\int_{\mathbf{Z}} \frac{\tilde{z}\phi(\sqrt{q^{(\ell)}}tr(\mathbf{\Sigma}^{(\ell)})/n\tilde{z})}{\sqrt{2\pi}} e^{-\frac{\tilde{z}^2}{2}} d\tilde{z}\right)^{2l}$$
$$\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.324}$$

$$= -\frac{\sigma_x^2\sigma_w^{2\ell}\left(\prod_{j=1}^{\ell}\theta_3^{(j)}\right)^2}{n^{1+l}}\left(\frac{1}{n}\sum_{k=1}^{n}\mathbf{\Sigma}_{kk}^{(\ell)l}\right)\left(\sqrt{q^{(\ell)}}\int_{\mathbf{Z}}\frac{\phi'(\sqrt{q^{(\ell)}}tr(\mathbf{\Sigma}^{(\ell)})/n\tilde{z})}{\sqrt{2\pi}}e^{-\frac{\tilde{z}^2}{2}}d\tilde{z}\right)^{2l}$$
$$\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.325}$$

$$= -\frac{\sigma_x^2\sigma_w^{2\ell}\left(\prod_{j=1}^{\ell}\theta_3^{(j)}\right)^2\tilde{\theta}_2^{(n)l}}{n^{1+l}}\left(\frac{1}{n}\sum_{k=1}^{n}\mathbf{\Sigma}_{kk}^{(\ell)l}\right)\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right). \tag{A.326}$$

□

A.3.2.1 Proof of Lemma 3.3.7

Proof. Lemma A.3.2 is now going to be used for the computation of E_{2k}. Assume that $k > 1$, and consider the contribution from an addend ω of (3.168) with n_ϕ and n_w factors of the type $\phi(\sum_q \mathbf{W}_{i_\xi,q}^{(\ell)}\mathbf{Y}_{q,\mu_\xi}^{(\ell)})\phi(\sum_q \mathbf{W}_{i_{\xi+1},q}^{(\ell)}\mathbf{Y}_{q,\mu_\xi}^{(\ell)})$ and $\sigma_x^2\left(\prod_{j=1}^{\ell}\theta_3^{(j)}\right)^2\mathfrak{W}_{i_\xi p}^{(\ell)}\mathfrak{W}_{i_{\xi+1}p}^{(\ell)}$ respectively, note that $k = n_w + n_\phi$. Then, given an index p

$$E_\omega^{(n,k,n_w,n_\phi,p)} = \int\prod_{\xi=1}^{n_w}\left(\sigma_x\left(\prod_{j=1}^{\ell}\theta_3^{(j)}\right)\mathfrak{W}_{i_\xi p}^{(\ell)}\left(\prod_{\ell=1}^{n_\phi^{(\xi)}}\phi(\sum_q \mathbf{W}_{i_{\xi+\ell},q}^{(\ell)}\mathbf{Y}_{q,\mu_{\xi+\ell}}^{(\ell)})\phi(\sum_q \mathbf{W}_{i_{\xi+\ell+1},q}^{(\ell)}Y_{q,\mu_{\xi+\ell}}^{(\ell)})\right)\right.$$
$$\left.\sigma_x\left(\prod_{j=1}^{\ell}\theta_3^{(j)}\right)\mathfrak{W}_{i_{\xi+1}p}^{(\ell)}\right)\mathcal{D}\mathbf{W}\mathcal{D}\mathbf{X}\mathcal{D}\mathbf{\Sigma}^{(\ell)} \tag{A.327}$$

$$= \int\prod_{\xi=1}^{n_w}E_{\Pi_\xi}^{(n,n_\phi^{(\xi)})}\mathcal{D}\mathbf{\Sigma}^{(\ell)} \tag{A.328}$$

$$= \int\prod_{\xi=1}^{n_w}-\frac{\sigma_w^{2\ell}\sigma_x^2\left(\prod_{j=1}^{\ell}\theta_3^{(j)}\right)^2\tilde{\theta}_2^{(n)n_\phi^{(\xi)}}}{n^{1+n_w^{(\xi)}}}\left(\frac{1}{n}\sum_{k=1}^{n}\mathbf{\Sigma}_{kk}^{(\ell)n_\phi^{(\xi)}}\right)\mathcal{D}\mathbf{\Sigma}^{(\ell)}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.329}$$

$$= \int\left(\frac{-\sigma_w^{2\ell}\sigma_x^2\left(\prod_{j=1}^{\ell}\theta_3^{(j)}\right)^2}{n}\right)^{n_w}\left(\frac{\tilde{\theta}_2^{(n)}}{n}\right)^{\sum_{\xi=1}^{n_w}n_\phi^{(\xi)}}\left(\prod_{\xi=1}\frac{1}{n}\sum_{k=1}^{n}\mathbf{\Sigma}_{kk}^{(\ell)n_\phi^{(\xi)}}\right)\mathcal{D}\mathbf{\Sigma}^{(\ell)}\left(1+\mathcal{O}\left(\frac{1}{n}\right)\right) \tag{A.330}$$

$$= \int \left(-\frac{\sigma_w^2 \sigma_x^2 \left(\prod_{j=1}^{\ell} \theta_3^{(j)} \right)^2}{n} \right)^{n_w} \left(\frac{\tilde{\theta}_2^{(n)}}{n} \right)^{k-n_w} \left(\prod_{\xi=1} \mu_{n_\phi^{(\xi)}} \right) \mathcal{D}\boldsymbol{\Sigma}^{(\ell)} \left(1 + \mathcal{O}\left(\frac{1}{n} \right) \right) \tag{A.331}$$

A.3.3 Lemma 3.3.8

Considering the covariance as $\boldsymbol{\Sigma} = \mathbf{I}$, then all the moments are equal to one. Therefore,

$$E_\omega^{(k,n_w,n_\phi,p)} = \left(-\frac{\sigma_w^2 \sigma_x^2 \left(\prod_{l=1}^{\ell} \theta_3^{(l)} \right)^2}{n} \right)^{n_w} \left(\frac{\theta_2}{n} \right)^{k-n_w} \left(1 + \mathcal{O}\left(\frac{1}{n} \right) \right). \tag{A.332}$$

Then the expected contribution for the $2k$ cycle is defined by considering the contribution $\tilde{E}_{2k}^{\phi} = n_0^{1-k} \theta_2^k \left(1 + \mathcal{O}\left(\frac{1}{n}\right)\right)$ of the addend ω whose $n_w = 0$ and of the addends for which $n_w \neq 0$

$$E_{2k} = \tilde{E}_{2k}^{\phi} + \sum_{n_w=1}^{k} \binom{k}{n_w} \sum_{p=1}^{n} E_\omega^{(k,n_w,n_\phi,p)} \tag{A.333}$$

$$= \left(n_0^{1-k} \theta_2^k + n_0^{1-k} \sum_{n_w=1}^{k} \binom{k}{n_w} \left(-\sigma_w^2 \sigma_x^2 \left(\prod_{l=1}^{\ell} \theta_3^{(l)} \right)^2 \right)^{n_w} \theta_2^{k-n_w} \right) \left(1 + \mathcal{O}\left(\frac{1}{n} \right) \right) \tag{A.334}$$

$$= \left(n_0^{1-k} \left(\theta_2 - \sigma_w^{2\ell} \sigma_x^2 \left(\prod_{l=1}^{\ell} \theta_3^{(l)} \right)^2 \right)^k \right) \left(1 + \mathcal{O}\left(\frac{1}{n} \right) \right) \tag{A.335}$$

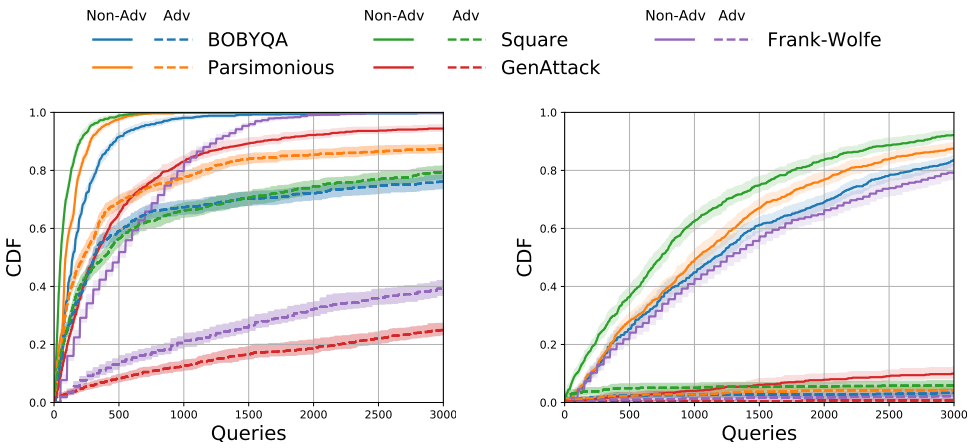
□

Appendix B

Experiments with Confidence Intervals

The results that are reported in Section 4.3 show how the different algorithms compare in expectation but do not give an estimate on how confident the results are. To get a measure of this, we bootstrap [22] the data. In particular, we do this by sampling 100 times 50% of the data with reinsertion for both the CIFAR10 and ImageNet case. In Figure B.1 and B.2 the 90% confidence interval is plotted in the shaded regions while the median result is plotted with the continuous line.

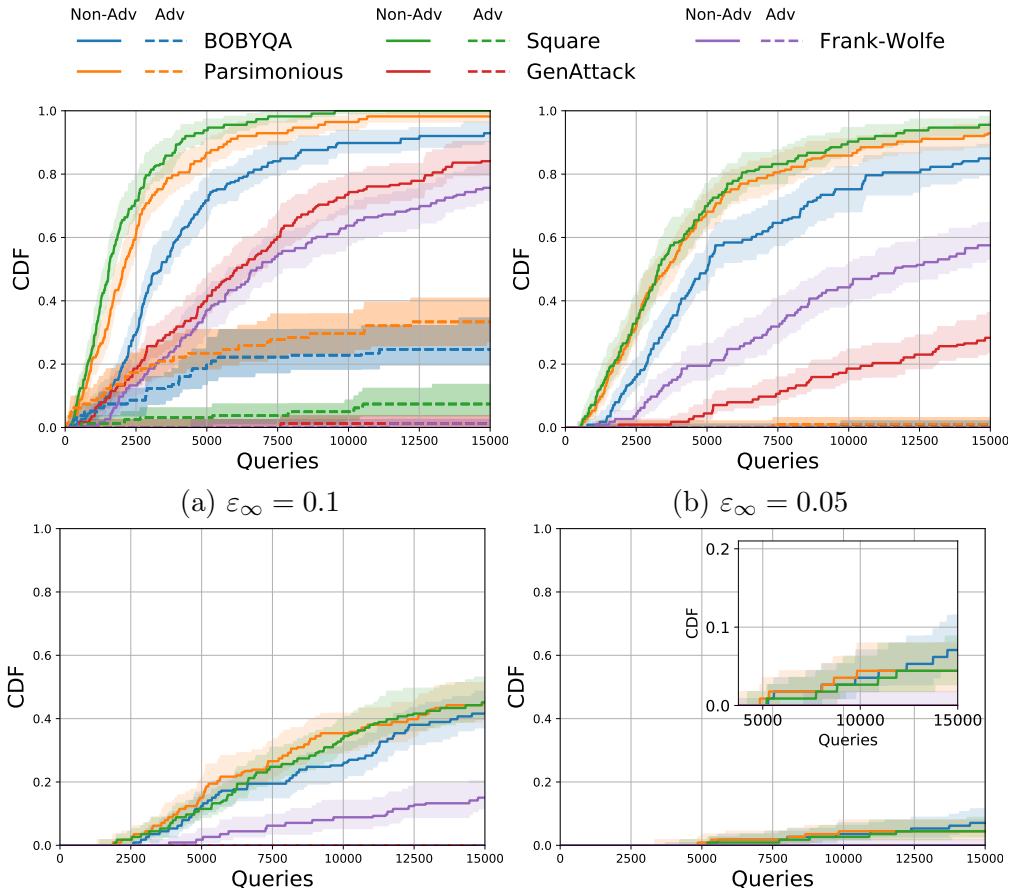
In the CIFAR10 case, the confidence regions are very near to the median values suggesting that the results that have been plotted in the main text have a very small variance. Since the shaded regions are almost distinct, the comparison between the methods is almost sure. On the other hand, the results relative to ImageNet have a larger confidence interval due to the fewer attacks considered. This is especially relevant for low energies as the confidence intervals overlap, and thus the median CDF of a method being higher than the other implies that one is more efficient than the other only with some probability. For example, for $\varepsilon = 0.01$ the BOBYQA algorithm is better than square with a confidence of 70%.



(a) $\varepsilon_\infty = 0.1$ (b) $\varepsilon_\infty = 0.02$

(c) $\varepsilon_\infty = 0.01$ (d) $\varepsilon_\infty = 0.005$

Figure B.1: Cumulative fraction of test set images successfully misclassified with adversarial examples generated by GenAttack, Parsimonious, Square, Frank-Wolfe, and our BOBYQA based approaches for different maximum perturbation energies ε_∞ and DNNs trained on the CIFAR10 dataset. In all results the solid and dashed lines denoted by 'Non-Adv' and 'Adv' corresponds to attacks on networks trained without or with the MadryLab defense strategy [24] respectively. The shaded region corresponds to the 90% confidence intervals for each method computed via boot-strapping the data.



(a) $\varepsilon_\infty = 0.1$ (b) $\varepsilon_\infty = 0.05$

(c) $\varepsilon_\infty = 0.02$ (d) $\varepsilon_\infty = 0.01$

Figure B.2: Cumulative fraction of test set images successfully misclassified with adversarial examples generated by GenAttack, Parsimonious, Square, Frank-Wolfe, and our BOBYQA based approaches for different maximum perturbation energies ε_∞ and DNNs trained on the ImageNet dataset. In all results the solid and dashed lines denoted by 'Non-Adv' and 'Adv' corresponds to attacks on networks trained without or with the MadryLab defense strategy [24] respectively. The shaded region corresponds to the 90% confidence intervals for each method computed via boot-strapping the data.

References

[1] Vinayak Abrol and Jared Tanner. Information-bottleneck under mean field initialization. *International Conference on Machine Learning (ICML) Workshop on Uncertainty and Robustness in Deep Learning*, 2020.

[2] Ben Adlam, Jake Levinson, and Jeffrey Pennington. A random matrix perspective on mixtures of nonlinearities in high dimensions, 2020. URL `https://openreview.net/forum?id=BJx7N1SKvB`.

[3] Abdullah Al-Dujaili and Una-May O'Reilly. There are no bit parts for sign bits in black-box attacks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2020.

[4] Moustafa Alzantot, Yash Sharma, Supriyo Chakraborty, Huan Zhang, Cho-Jui Hsieh, and Mani B. Srivastava. Genattack: Practical black-box attacks with gradient-free optimization. In *Proceedings of the Genetic and Evolutionary Computation Conference (GECCO)*, page 1111–1119, July 2019. doi: 10.1145/3321707.3321749.

[5] Maksym Andriushchenko, Francesco Croce, Nicolas Flammarion, and Matthias Hein. Square attack: a query-efficient black-box adversarial attack via random search. *Proceedings of the European Conference on Computer Vision (ECCV)*, 2020.

[6] Lucas Benigni and Sandrine Péché. Eigenvalue distribution of some nonlinear models of random matrices. *Electronic Journal of Probability*, 26:1–37, 2021.

[7] Albert S Berahas, Majid Jahani, and Martin Takác. Quasi-newton methods for deep learning: Forget the past, just sample. *arXiv preprint arXiv:1901.09997*, 16, 2019.

[8] Léon Bottou. Large-scale machine learning with stochastic gradient descent. In *Proceedings of COMPSTAT'2010*, pages 177–186. Springer, 2010.

[9] N. Carlini and D. Wagner. Towards evaluating the robustness of neural networks. In *Proceedings of the IEEE Symposium on Security and Privacy (SP)*, pages 39–57, May 2017. doi: 10.1109/SP.2017.49.

[10] Coralia Cartis, Jan Fiala, Benjamin Marteau, and Lindon Roberts. Improving the flexibility and robustness of model-based derivative-free optimization solvers. *ACM Trans. Math. Softw.*, 45(3), August 2019. doi: 10.1145/3338517.

[11] Coralia Cartis, Tyler Ferguson, and Lindon Roberts. Scalable derivative-free optimization for nonlinear least-square problems. *Beyond First-Order Methods in ML Systems workshop at ICML*, 2020.

[12] Jinghui Chen, Dongruo Zhou, Jinfeng Yi, and Quanquan Gu. A Frank-Wolfe framework for efficient and effective adversarial attacks. In *Proceedings of the Association for the Advancement of Artificial Intelligence Conference (AAAI)*, pages 3486–3494, 2020.

[13] Pin-Yu Chen, Huan Zhang, Yash Sharma, Jinfeng Yi, and Cho-Jui Hsieh. ZOO: zeroth order optimization based black-box attacks to deep neural networks without training substitute models. In *Proceedings of the ACM Workshop on Artificial Intelligence and Security (AISec)*, page 15–26, November 2017. doi: 10.1145/3128572.3140448.

[14] Pin-Yu Chen, Yash Sharma, Huan Zhang, Jinfeng Yi, and Cho-Jui Hsieh. Ead: elastic-net attacks to deep neural networks via adversarial examples. In *Proceedings of the AAAI Conference on Artificial Intelligence*, pages 10–17, February 2018.

[15] Bobby Chesney and Danielle Citron. Deep fakes: A looming challenge for privacy, democracy, and national security. *Calif. L. Rev.*, 107:1753, 2019.

[16] Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *Artificial intelligence and statistics*, pages 192–204. PMLR, 2015.

[17] Andrew R Conn, Katya Scheinberg, and Luis N Vicente. *Introduction to derivative-free optimization*, volume 8. SIAM, 2009.

[18] George Cybenko. Approximation by superpositions of a sigmoidal function. *Mathematics of control, signals and systems*, 2(4):303–314, 1989.

[19] Nilesh Dalvi, Pedro Domingos, Sumit Sanghai, Deepak Verma, et al. Adversarial classification. In *Proceedings of the ACM International conference on Knowledge Discovery and Data Mining (SIGKDD)*, pages 99–108, August 2004. doi: 10.1145/1014052.1014066.

[20] Guneet S Dhillon, Kamyar Azizzadenesheli, Zachary C Lipton, Jeremy Bernstein, Jean Kossaifi, Aran Khanna, and Anima Anandkumar. Stochastic activation pruning for robust adversarial defense. In *Proceedings of the International Conference on Learning Representations (ICLR)*, April 2018.

[21] Alan Edelman and N Raj Rao. Random matrix theory. *Acta numerica*, 14: 233–297, 2005.

[22] Bradley Efron and Robert J Tibshirani. *An introduction to the bootstrap*. CRC press, 1994.

[23] Fidah El Haje Hussein and Yurii Golubev. On entropy estimation by m-spacing method. *Записки научных семинаров ПОМИ*, 363(0):151–181, 2009.

[24] Logan Engstrom, Andrew Ilyas, Shibani Santurkar, and Dimitris Tsipras. Robustness (python library), 2019. URL `https://github.com/MadryLab/robustness`.

[25] K. Eykholt, I. Evtimov, E. Fernandes, B. Li, A. Rahmati, C. Xiao, A. Prakash, T. Kohno, and D. Song. Robust physical-world attacks on deep learning visual classification. In *Proceedings of the IEEE/CVF Conference on Computer Vision and Pattern Recognition (CVPR)*, pages 1625–1634, June 2018. doi: 10.1109/CVPR.2018.00175.

[26] Brandon Foggo and Nanpeng Yu. On the maximum mutual information capacity of neural architectures. *arXiv preprint arXiv:2006.06037*, 2020.

[27] GW Ford and GE Uhlenbeck. Combinatorial problems in the theory of graphs. iii. *Proceedings of the National Academy of Sciences of the United States of America*, 42(8):529, 1956.

[28] Marylou Gabrié, Andre Manoel, Clément Luneau, Jean Barbier, Nicolas Macris, Florent Krzakala, and Lenka Zdeborová. Entropy and mutual information in models of deep neural networks. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124014, 2019.

[29] Xavier Glorot and Yoshua Bengio. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the thirteenth international conference on artificial intelligence and statistics*, pages 249–256. JMLR Workshop and Conference Proceedings, 2010.

[30] Ziv Goldfeld, Ewout van den Berg, Kristjan Greenewald, Igor Melnyk, Nam Nguyen, Brian Kingsbury, and Yury Polyanskiy. Estimating information flow in deep neural networks. *International Conference on Machine Learning (ICML)*, 2018.

[31] Ian Goodfellow, Yoshua Bengio, and Aaron Courville. *Deep Learning*. MIT Press, 2016. `http://www.deeplearningbook.org`.

[32] Ian J Goodfellow, Jonathon Shlens, and Christian Szegedy. Explaining and harnessing adversarial examples. In *Proceedings of the International Conference on Learning Representations (ICLR)*, May 2015.

[33] Soorya Gopalakrishnan, Zhinus Marzi, Upamanyu Madhow, and Ramtin Pedarsani. Toward robust neural networks via sparsification. *arXiv preprint arXiv:1810.10625*, 2018.

[34] Serge Gratton, Annick Sartenaer, and Philippe L Toint. Recursive trust-region methods for multiscale nonlinear optimization. *SIAM Journal on Optimization*, 19(1):414–444, 2008.

[35] Chuan Guo, Jared S Frank, and Kilian Q Weinberger. Low frequency adversarial perturbation. In *Proceedings of the Conference on Uncertainty in Artificial Intelligence (UAI)*, July 2018.

[36] Chuan Guo, Mayank Rana, Moustapha Cisse, and Laurens van der Maaten. Countering adversarial images using input transformations. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2018. URL `https://openreview.net/forum?id=SyJ7ClWCb`.

[37] Han Xu Yao Ma Hao-Chen, Liu Debayan Deb, Hui Liu Ji-Liang Tang Anil, and K Jain. Adversarial attacks and defenses in images, graphs and text: A review. *International Journal of Automation and Computing*, 17(2):151–178, 2020.

[38] Kaiming He, Xiangyu Zhang, Shaoqing Ren, and Jian Sun. Deep residual learning for image recognition. In *Proceedings of the IEEE conference on computer vision and pattern recognition*, pages 770–778, 2016.

[39] Kurt Hornik, Maxwell Stinchcombe, and Halbert White. Multilayer feedforward networks are universal approximators. *Neural networks*, 2(5):359–366, 1989.

[40] Stefan Höst. *Information and Communication Theory*. Wiley, 9 2019. ISBN 9781119433781. doi: 10.1002/9781119433828.

[41] Andrew Ilyas, Logan Engstrom, Anish Athalye, and Jessy Lin. Black-box adversarial attacks with limited queries and information. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 2137–2146, July 2018.

[42] Andrew Ilyas, Logan Engstrom, and Aleksander Madry. Prior convictions: Black-box adversarial attacks with bandits and priors. In *Proceedings of the International Conference on Learning Representations (ICLR)*, 2019.

[43] Leon Isserlis. On a formula for the product-moment coefficient of any order of a normal frequency distribution in any number of variables. *Biometrika*, 12(1/2):134–139, 1918.

[44] Jean Jacod and Philip Protter. *Probability essentials*. Springer Science & Business Media, 2004.

[45] Iain M Johnstone. High dimensional statistical inference and random matrices. *arXiv preprint math/0611589*, 2006.

[46] John Jumper, Richard Evans, Alexander Pritzel, Tim Green, Michael Figurnov, Olaf Ronneberger, Kathryn Tunyasuvunakool, Russ Bates, Augustin Žídek, Anna Potapenko, et al. Highly accurate protein structure prediction with alphafold. *Nature*, 596(7873):583–589, 2021.

[47] Jon Keating. Lecture notes in random matrix theory, February 2021.

[48] Artemy Kolchinsky, Brendan D Tracey, and Steven Van Kuyk. Caveats for information bottleneck in deterministic scenarios. *International Conference on Learning Representations (ICLR)*, 2019.

[49] Alexey Kurakin, Ian Goodfellow, and Samy Bengio. Adversarial examples in the physical world. In *Proceedings of the International Conference on Learning Representations (ICLR), Workshop Track*, April 2017.

[50] Jeffrey Larson, Matt Menickelly, and Stefan M. Wild. Derivative-free optimization methods. *Acta Numerica*, 28:287–404, 2019. doi: 10.1017/S0962492919000060.

[51] Yann LeCun, Yoshua Bengio, and Geoffrey Hinton. Deep learning. *Nature*, 521(7553):436–444, 2015.

[52] R. Linsker. Self-organization in a perceptual network. *Computer*, 21(3): 105–117, 1988. doi: 10.1109/2.36.

[53] Yanpei Liu, Xinyun Chen, Chang Liu, and Dawn Song. Delving into transferable adversarial examples and black-box attacks. *Proceedings of the International Conference on Learning Representations (ICLR)*, 2017.

[54] Vladimir A Marčenko and Leonid Andreevich Pastur. Distribution of eigenvalues for some sets of random matrices. *Mathematics of the USSR-Sbornik*, 1(4):457, 1967.

[55] Arnaud Marsiglietti and Victoria Kostina. A lower bound on the differential entropy of log-concave random vectors with applications. *Entropy*, 20(3): 185, 2018.

[56] James Martens and Roger Grosse. Optimizing neural networks with kronecker-factored approximate curvature. In *International conference on machine learning*, pages 2408–2417. PMLR, 2015.

[57] Robert M May. Will a large complex system be stable? *Nature*, 238(5364): 413–414, 1972.

[58] David McAllester and Karl Stratos. Formal limitations on the measurement of mutual information. In *International Conference on Artificial Intelligence and Statistics*, pages 875–884. PMLR, 2020.

[59] James A Mingo and Roland Speicher. *Free probability and random matrices*, volume 35. Springer, 2017.

[60] Seungyong Moon, Gaon An, and Hyun Oh Song. Parsimonious black-box adversarial attacks via efficient combinatorial optimization. In *Proceedings of the International Conference on Machine Learning (ICML)*, pages 4636–4645, June 2019.

[61] Jorge J Moré and Stefan M Wild. Benchmarking derivative-free optimization algorithms. *SIAM Journal on Optimization*, 20(1):172–191, 2009.

[62] Kevin P Murphy. *Machine learning: a probabilistic perspective*. MIT press, 2012.

[63] N. Narodytska and S. Kasiviswanathan. Simple black-box adversarial attacks on deep neural networks. In *Proceedings of the IEEE Conference on Computer Vision and Pattern Recognition Workshops (CVPRW)*, pages 1310–1318, July 2017. doi: 10.1109/CVPRW.2017.172.

[64] M. Naseer, S. Khan, and F. Porikli. Local gradients smoothing: Defense against localized adversarial attacks. In *2019 IEEE Winter Conference on Applications of Computer Vision (WACV)*, pages 1300–1307, 2019.

[65] Jorge Nocedal and Stephen J. Wright. *Numerical Optimization*. Springer-Verlag New York, 2006. ISBN 978-0-387-30303-1. doi: 10.1007/978-0-387-40065-5.

[66] Morteza Noshad, Yu Zeng, and Alfred O Hero. Scalable mutual information estimation using dependence graphs. In *ICASSP 2019-2019 IEEE International Conference on Acoustics, Speech and Signal Processing (ICASSP)*, pages 2962–2966. IEEE, 2019.

[67] Odhran O'Donoghue, Paul Duckworth, Giuseppe Ughi, Linus Scheibenreif, Kia Khezeli, Adrienne Hoarfrost, Samuel Budd, Patrick Foley, Nicholas Chia, John Kalantari, et al. Invariant risk minimisation for cross-organism inference: Substituting mouse data for human data in human risk factor discovery. *Proceedings of the Neural Information Processing Systems (NeruIPS) workshop "Machine Learning for Health"*, 2021.

[68] Aaron van den Oord, Yazhe Li, and Oriol Vinyals. Representation learning with contrastive predictive coding. *ArXiv preprint arXiv:1807.03748*, 2018.

[69] Liam Paninski. Estimation of entropy and mutual information. *Neural computation*, 15(6):1191–1253, 2003.

[70] Nicolas Papernot, Patrick McDaniel, Ian Goodfellow, Somesh Jha, Z Berkay Celik, and Ananthram Swami. Practical black-box attacks against machine learning. In *Proceedings of the 2017 ACM on Asia Conference on Computer and Communications Security*, pages 506–519. ACM, 2017.

[71] Jeffrey Pennington and Yasaman Bahri. Geometry of neural network loss surfaces via random matrix theory. In *Proceedings of the 34th International Conference on Machine Learning-Volume 70*, pages 2798–2806. JMLR. org, 2017.

[72] Jeffrey Pennington and Pratik Worah. Nonlinear random matrix theory for deep learning. *Neural Information Processing Systems (NeurIPS)*, 2017.

[73] Jeffrey Pennington, Samuel S Schoenholz, and Surya Ganguli. Resurrecting the sigmoid in deep learning through dynamical isometry: theory and practice. In *Proceedings of the 31st International Conference on Neural Information Processing Systems*, pages 4788–4798, 2017.

[74] Jeffrey Pennington, Samuel Schoenholz, and Surya Ganguli. The emergence of spectral universality in deep networks. *Artificial Intelligence and Statistics (AISTATS)*, 2 2018.

[75] B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli. Exponential expressivity in deep neural networks through transient chaos. *Neural Information Processing Systems (NeurIPS)*, 2016.

[76] Ben Poole, Subhaneil Lahiri, Maithra Raghu, Jascha Sohl-Dickstein, and Surya Ganguli. Exponential expressivity in deep neural networks through transient chaos. In *Advances in neural information processing systems*, pages 3360–3368, 2016.

[77] Ben Poole, Sherjil Ozair, Aaron Van Den Oord, Alex Alemi, and George Tucker. On variational bounds of mutual information. In *International Conference on Machine Learning*, pages 5171–5180. PMLR, 2019.

[78] Martin Popel, Marketa Tomkova, Jakub Tomek, Łukasz Kaiser, Jakob Uszkoreit, Ondřej Bojar, and Zdeněk Žabokrtskỳ. Transforming machine translation: a deep learning system reaches news translation quality comparable to human professionals. *Nature communications*, 11(1):1–15, 2020.

[79] Marc Potters and Jean-Philippe Bouchaud. *A First Course in Random Matrix Theory: For Physicists, Engineers and Data Scientists*. Cambridge University Press, 2020.

[80] Michael JD Powell. The BOBYQA algorithm for bound constrained optimization without derivatives. Technical Report DAMTP 2009/NA06, University of Cambridge, 2009.

[81] Kui Ren, Tianhang Zheng, Zhan Qin, and Xue Liu. Adversarial attacks and defenses in deep learning. *Engineering*, 6(3):346–360, 2020.

[82] Lindon Roberts. *Derivative-free algorithms for nonlinear optimisation problems*. PhD thesis, University of Oxford, 2019.

[83] David E Rumelhart, Geoffrey E Hinton, and Ronald J Williams. Learning representations by back-propagating errors. *nature*, 323(6088):533–536, 1986.

[84] Olga Russakovsky, Jia Deng, Hao Su, Jonathan Krause, Sanjeev Satheesh, Sean Ma, Zhiheng Huang, Andrej Karpathy, Aditya Khosla, Michael Bernstein, et al. Imagenet large scale visual recognition challenge. *International journal of computer vision*, 115(3):211–252, 2015.

[85] Andrew M Saxe, Yamini Bansal, Joel Dapello, Madhu Advani, Artemy Kolchinsky, Brendan D Tracey, and David D Cox. On the information bottleneck theory of deep learning. *Journal of Statistical Mechanics: Theory and Experiment*, 2019(12):124020, 2019.

[86] K. Scheinberg and P. Toint. Self-correcting geometry in model-based algorithms for derivative-free unconstrained optimization. *SIAM J. Optim.*, 20: 3512–3532, 2010.

[87] Konrad Schmüdgen. *The moment problem*, volume 9. Springer, 2017.

[88] Samuel S Schoenholz, Justin Gilmer, Surya Ganguli, and Jascha Sohl-Dickstein. Deep information propagation. *International Conference on Learning Representations (ICLR)*, 2017.

[89] Alexandru Constantin Serban and Erik Poll. Adversarial examples-a complete characterisation of the phenomenon. *arXiv preprint arXiv:1810.01185*, 2018.

[90] Yash Sharma, Gavin Weiguang Ding, and Marcus Brubaker. On the effectiveness of low frequency perturbations. *Proceedings of the International Joint Conference on Artificial Intelligence (IJCAI)*, pages 3389–3396, August 2019. doi: 10.24963/ijcai.2019/470.

[91] Ravid Shwartz-Ziv and Naftali Tishby. Opening the black box of deep neural networks via information. 2017.

[92] David Silver, Julian Schrittwieser, Karen Simonyan, Ioannis Antonoglou, Aja Huang, Arthur Guez, Thomas Hubert, Lucas Baker, Matthew Lai, Adrian Bolton, et al. Mastering the game of go without human knowledge. *Nature*, 550(7676):354–359, 2017.

[93] Karen Simonyan and Andrew Zisserman. Very deep convolutional networks for large-scale image recognition. In *International Conference on Learning Representations*, 2015.

[94] Chawin Sitawarin, Arjun Nitin Bhagoji, Arsalan Mosenia, Mung Chiang, and Prateek Mittal. Darts: Deceiving autonomous cars with toxic signs. *arXiv preprint arXiv:1802.06430*, 2018.

[95] Roland Speicher. Free probability theory. *Jahresbericht der Deutschen Mathematiker-Vereinigung*, 119(1):3–30, 2017.

[96] Joel Spencer. *Asymptopia*, volume 71. American Mathematical Soc., 2014.

[97] Gilbert Strang. *Introduction to linear algebra*, volume 3. Wellesley-Cambridge Press Wellesley, MA, 1993.

[98] Christian Szegedy, Wojciech Zaremba, Ilya Sutskever, Joan Bruna, Dumitru Erhan, Ian Goodfellow, and Rob Fergus. Intriguing properties of neural networks. In *Proceedings of the International Conference on Learning Representations (ICLR)*, April 2014.

[99] J Tanner and G Ughi. Mutual information of neural network initialisations: mean field approximations. In *Proceedings of the IEEE International Symposium on Information Theory*. IEEE, 2021.

[100] Matus Telgarsky. Representation benefits of deep feedforward networks. *arXiv preprint arXiv:1509.08101*, 2015.

[101] Simon F. B. Tett, Michael J. Mineter, Coralia Cartis, Daniel J. Rowlands, and Ping Liu. Can top-of-atmosphere radiation measurements constrain climate predictions? part i: Tuning. *Journal of Climate*, 26(23):9348–9366, November 2013. doi: 10.1175/JCLI-D-12-00595.1.

[102] Chun-Chen Tu, Paishun Ting, Pin-Yu Chen, Sijia Liu, Huan Zhang, Jinfeng Yi, Cho-Jui Hsieh, and Shin-Ming Cheng. Autozoom: Autoencoder-based zeroth order optimization method for attacking black-box neural networks. In *Proceedings of the AAAI Conference on Artificial Intelligence: Special Technical Track: AI for Social Impact*, July 2019. doi: 10.1609/aaai.v33i01.3301742.

[103] Giuseppe Ughi, Vinayak Abrol, and Jared Tanner. A model-based derivative-free approach to black-box adversarial examples: BOBYQA. *Proceedings of the Neural Information Processing Systems (NeruIPS) workshop "Beyond First Order Methods in ML"*, 2019.

[104] Giuseppe Ughi, Vinayak Abrol, and Jared Tanner. An empirical study of derivative-free-optimization algorithms for targeted black-box attacks in deep neural networks. *Optimization and Engineering*, pages 1–28, 2021.

[105] Oriol Vinyals, Igor Babuschkin, Wojciech M Czarnecki, Michaël Mathieu, Andrew Dudzik, Junyoung Chung, David H Choi, Richard Powell, Timo Ewalds, Petko Georgiev, et al. Grandmaster level in StarCraft II using multi-agent reinforcement learning. *Nature*, 575(7782):350–354, 2019.

[106] Xiao Wang, Siyue Wang, Pin-Yu Chen, Yanzhi Wang, Brian Kulis, Xue Lin, and Sang Chin. Protecting neural networks with hierarchical random switching: Towards better robustness-accuracy trade-off for stochastic defenses. In

Proceedings of the International Joint Conference on Artificial Intelligence, IJCAI, August 2019.

[107] Andreas Winkelbauer. Moments and absolute moments of the normal distribution. *CoRR*, abs/1209.4340, 2012. URL `http://arxiv.org/abs/1209.4340`.

[108] Huikai Wu, Shuai Zheng, Junge Zhang, and Kaiqi Huang. GP-Gan: Towards realistic high-resolution image blending. In *Proceedings of the 27th ACM international conference on multimedia*, pages 2487–2495, 2019.

[109] Jiajun Wu, Chengkai Zhang, Tianfan Xue, William T Freeman, and Joshua B Tenenbaum. Learning a probabilistic latent space of object shapes via 3d generative-adversarial modeling. In *Proceedings of the 30th International Conference on Neural Information Processing Systems*, pages 82–90, 2016.

[110] Dmitry Yarotsky. Error bounds for approximations with deep relu networks. *Neural Networks*, 94:103–114, 2017.

[111] X. Yuan, P. He, Q. Zhu, and X. Li. Adversarial examples: Attacks and defenses for deep learning. *IEEE Transactions on Neural Networks and Learning Systems*, 30(9):2805–2824, September 2019. doi: 10.1109/TNNLS.2018.2886017.

[112] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In *International Conference on Machine Learning (ICML)*, pages 7472–7482. PMLR, 2019.

[113] Hongyang Zhang, Yaodong Yu, Jiantao Jiao, Eric Xing, Laurent El Ghaoui, and Michael Jordan. Theoretically principled trade-off between robustness and accuracy. In Kamalika Chaudhuri and Ruslan Salakhutdinov, editors, *Proceedings of the 36th International Conference on Machine Learning*, volume 97 of *Proceedings of Machine Learning Research*, pages 7472–7482. PMLR, 09–15 Jun 2019. URL `https://proceedings.mlr.press/v97/zhang19p.html`.