Nonlinear random matrix theory for deep learning

Jeffrey Pennington
Google Brain
jpennin@google.com

Pratik Worah
Google Research
pworah@google.com

Abstract

Neural network configurations with random weights play an important role in the analysis of deep learning. They define the initial loss landscape and are closely related to kernel and random feature methods. Despite the fact that these networks are built out of random matrices, the vast and powerful machinery of random matrix theory has so far found limited success in studying them. A main obstacle in this direction is that neural networks are nonlinear, which prevents the straightforward utilization of many of the existing mathematical results. In this work, we open the door for direct applications of random matrix theory to deep learning by demonstrating that the pointwise non-linearities typically applied in neural networks can be incorporated into a standard method of proof in random matrix theory known as the moments method. The test case for our study is the Gram matrix Y^TY, $Y = f(WX)$, where W is a random weight matrix, X is a random data matrix, and f is a point-wise non-linear activation function. We derive an explicit representation for the trace of the resolvent of this matrix, which defines its limiting spectral distribution. We apply these results to the computation of the asymptotic performance of single-layer random feature networks on a memorization task and to the analysis of the eigenvalues of the data covariance matrix as it propagates through a neural network. As a byproduct of our analysis, we identify an intriguing new class of activation functions with favorable properties.

Keywords: Neural networks, Random Matrix Theory, Moment method.

1. Introduction

The list of successful applications of deep learning is growing at a staggering rate. It has been successfully applied to classical problems in Image recognition (Krizhevsky et al., 2012), audio synthesis (Oord et al., 2016), translation (Wu et al., 2016), speech recognition (Hinton et al., 2012), and there have been several applications to unexpected areas including protein structure prediction (Goh et al., 2017), quantum chemistry (Goh et al., 2017) and drug discovery (Altae-Tran et al., 2017). Our theoretical understanding of deep learning, on the other hand, has progressed at a more modest pace. A central difficulty in extending our understanding stems from the complexity of neural network loss surfaces, which are highly non-convex functions, often of millions or even billions (Shazeer et al., 2017) of parameters.

In the physical sciences, progress in understanding large complex systems has often come by approximating their constituents with random variables; for example, statistical physics and thermodynamics are based in this paradigm. Since modern neural networks are undeniably large complex systems, it is natural to consider what insights can be gained by approximating their parameters with random variables. Moreover, such random configurations play at least two privileged roles in neural networks: they define the initial loss surface for optimization, and they are closely related to random feature and kernel methods. Therefore it is not surprising that random neural networks have attracted significant attention in the literature in recent years.

Another useful technique for simplifying the study of large complex systems is to approximate their size as infinite. For neural networks, the concept of size has at least two axes: the number of samples and the number of parameters. It is common, particularly in the statistics literature, to consider the mean performance of a finite-capacity model against a given data distribution. From this perspective, the number of samples, m, is taken to be infinite relative to the number of parameters, n, i.e. $n/m \to 0$. An alternative perspective is employed in the study of kernel or random feature methods. In this case, the number of parameters is frequently taken to be infinite relative to the number of samples, i.e. $n/m \to \infty$. In practice, however, most successful modern deep learning architectures tend to have both a large number of samples *and* a large number of parameters, often of roughly the same order of magnitude. One simple explanation for this may just be that the other extremes tend to produce over- or under-fitting. Motivated by this observation, in this work we explore the infinite size limit in which both the number of samples and the number of parameters go to infinity at the same rate, i.e. $n/m \to \phi$, for some finite constant ϕ. This perspective puts us squarely in the regime of random matrix theory.

An abundance of matrices are of practical and theoretical interest in the context of random neural networks. For example, the input-output Jacobian, the Hessian of the loss function with respect to the weights, or simply the output of the network are all interesting objects of study. In this work we focus on the computation of the eigenvalues of the matrix $M \equiv \frac{1}{m} Y^T Y$, where $Y = f(WX)$, W is a Gaussian random weight matrix, X is a Gaussian random data matrix, and f is a pointwise activation function. In many ways, Y is a basic primitive whose understanding is necessary for attacking more complicated cases; for example, Y shows up in the expressions for all three of the matrices mentioned above. But studying Y is also quite interesting in its own right, with several interesting applications to machine learning that we will explore in Section 5.

1.1 Our contribution

The non-linearity of the activation function prevents us from leveraging many of the existing mathematical results from random matrix theory. Nevertheless, most of the basic tools for computing spectral densities of random matrices still apply in this setting. In this work, we show how to overcome some of the technical hurdles that have prevented explicit computations of this type in the past. In particular, we employ the so-called *moments method*, deducing the spectral density of M from the traces $\operatorname{tr} M^k$. Evaluating the traces involves computing certain multi-dimensional integrals, which we show how to evaluate, and

enumerating a certain class of graphs, for which we derive a generating function. The result of our calculation is a quartic equation which is satisfied by the trace of the resolvent of M, $G(z) = -\mathbf{E}\,\mathrm{tr}(M - zI)^{-1}$. The techniques presented here pave the way for studying other types of non-linear random matrices relevant for the theoretical understanding of neural networks.

1.2 Applications of our results

We show that the training loss of a ridge-regularized single-layer random-feature least-squares memorization problem with regularization parameter γ is related to $-\gamma^2 G'(-\gamma)$. As such, we obtain an expression for the training loss as a function of γ, and observe its dependence on two parameters η and ζ, defined in Theorem 1, that capture the (only) relevant properties of the non-linearity f. We observe increased memorization capacity for certain types of non-linearities relative to others. In particular, for a fixed value of γ, the training loss is lower if η/ζ is large, a condition satisfied if f is close to an even function with zero Gaussian mean. We believe this observation could have an important practical impact in designing next-generation activation functions.

We also examine the eigenvalue density of M and observe that if $\zeta = 0$ the distribution collapses to the Marchenko-Pastur distribution (Marčenko and Pastur, 1967), which describes the eigenvalues of the Wishart matrix $X^T X$. We therefore make the surprising observation that there exist functions f such that $f(WX)$ has the same singular value distribution as X. Said another way, the eigenvalues of the data covariance matrix are unchanged in distribution after passing through a single non-linear layer of the network. We conjecture that this property is actually satisfied through arbitrary layers of the network, and find supporting numerical evidence. This conjecture may be regarded as a claim on the *universality* of our results with respect to the distribution of X. Note that preserving the first moment of this distribution is also an effect achieved through batch normalization (Ioffe and Szegedy, 2015), although higher moments are not necessarily preserved. We therefore offer the hypothesis that choosing activation functions with $\zeta = 0$ might lead to improved training performance, in the same way that batch normalization does, at least early in training.

1.3 Related work

Random neural networks have been studied from a variety of angles. There has been much experimentation since the 90s to study the convergence properties and therefore the structure of loss functions of neural networks. Pierre and Kurt (1989) analyzed the error surface of a multilayer perceptron with a single hidden layer. Saad and Solla (1995) analyzed the dynamics of stochastic gradient descent for soft committee machines using methods from statistical physics. The edited work (Saad, 2009) compiles several references from the previous couple of decades that use statistical physics methods for the analysis of gradient based methods in deep neural nets.

More recently, Bray and Dean (2007) and Fyodorov and Williams (2007) study the nature of critical points of random Gaussian error functions using replica theory. In the light of random matrix theory their results may be roughly interpreted as an increase in density of eigenvalues around zero. Even more recently, Choromanska et al. (2015) and Dauphin et al. (2014) studied the critical points of random loss surfaces and identified the prevalence

of saddle points, their work establishes a strong connection between random Gaussian fields and loss surfaces; Saxe et al. (2014) examined the effect of random initialization on the dynamics of learning in deep linear networks; Schoenholz et al. (2016) studied how information propagates through random networks, and how that affects learning; Poole et al. (2016) and Raghu et al. (2016) investigated various measures of expressivity in the context of deep random neural networks. Approximate kernel and random feature methods provide another application of random networks (Rahimi and Recht, 2007; Daniely et al., 2016; Neal, 2012).

However, so far there have not been many attempts to study random matrices with non-linear dependencies. The little work in this direction has focused on kernel random matrices and robust statistics models (El Karoui et al., 2010; Cheng and Singer, 2013). In a closely related recent work, Louart et al. (2017) examined the resolvent of Gram matrix YY^T in the case where X is deterministic.

2. Preliminaries

Throughout this work we will be relying on a number of basic concepts from random matrix theory. Here we provide a lightning overview of the essentials, but refer the reader to the more pedagogical literature for background (Tao, 2012).

2.1 Notation

Let $X \in \mathbb{R}^{n_0 \times m}$ be a random data matrix with i.i.d. elements $X_{i\mu} \sim \mathcal{N}(0, \sigma_x^2)$ and $W \in \mathbb{R}^{n_1 \times n_0}$ be a random weight matrix with i.i.d. elements $W_{ij} \sim \mathcal{N}(0, \sigma_w^2/n_0)$. As discussed in Section 1, we are interested in the regime in which both the row and column dimensions of these matrices are large and approach infinity at the same rate. In particular, we define

$$\phi \equiv \frac{n_0}{m}, \quad \psi \equiv \frac{n_0}{n_1}\,, \tag{1}$$

to be fixed constants as $n_0, n_1, m \to \infty$.

We denote the matrix of pre-activations by $Z = WX$. Let $f : \mathcal{R} \to \mathcal{R}$ be a function with zero mean and finite moments,

$$\int \frac{dz}{\sqrt{2\pi}}\, e^{-\frac{z^2}{2}} f(\sigma_w \sigma_x z) = 0, \qquad \left| \int \frac{dz}{\sqrt{2\pi}}\, e^{-\frac{z^2}{2}} f(\sigma_w \sigma_x z)^k \right| < \infty \;\text{ for }\; k > 1\,, \tag{2}$$

and denote the matrix of post-activations $Y = f(Z)$, where f is applied point-wise. We will be interested in the Gram matrix,

$$M = \frac{1}{m} YY^T \in \mathbb{R}^{n_1 \times n_1}\,. \tag{3}$$

2.2 Spectral density and the Stieltjes transform

The *empirical spectral density* of M is defined as,

$$\rho_M = \frac{1}{n_1} \sum_{j=1}^{n_1} \delta_{\lambda_j(M)}\,, \tag{4}$$

where the $\lambda_j(M)$, $j = 1, \ldots, n_1$, denote the n_1 eigenvalues of M, including multiplicity. The *limiting spectral density* is defined as the limit of Equation 4 as $n_1 \to \infty$, if it exists.

For $z \in \mathbb{C} \setminus \mathrm{supp}(\rho_M)$ the *Stieltjes transform* G of ρ_M is defined as,

$$G(z) = \int \frac{\rho_M(t)}{z - t} dt = -\frac{1}{n_1} \mathbf{E} \operatorname{tr}(M - zI_{n_1})^{-1} , \tag{5}$$

where the expectation is with respect to the random variables W and X. The quantity $(M - zI_{n_1})^{-1}$ is the *resolvent* of M. The spectral density can be recovered from the Stieltjes transform using the inversion formula,

$$\rho_M(\lambda) = -\frac{1}{\pi} \lim_{\epsilon \to 0^+} \operatorname{Im} G(\lambda + i\epsilon) . \tag{6}$$

2.3 Moment method

One of the main tools for computing the limiting spectral distributions of random matrices is the moment method, which, as the name suggests, is based on computations of the moments of ρ_M. The asymptotic expansion of Equation 5 for large z gives the Laurent series,

$$G(z) = \sum_{k=0}^{\infty} \frac{m_k}{z^{k+1}} , \tag{7}$$

where m_k is the kth moment of the distribution ρ_M,

$$m_k = \int dt \; \rho_M(t) t^k = \frac{1}{n_1} \mathbf{E} \operatorname{tr} M^k . \tag{8}$$

If one can compute m_k, then the density ρ_M can be obtained via eqns. (7) and (6). The idea behind the moment method is to compute m_k by expanding out powers of M inside the trace as,

$$\frac{1}{n_1} \mathbf{E} \operatorname{tr} M^k = \frac{1}{n_1} \mathbf{E} \sum_{i_1, \ldots, i_k \in [n_1]} M_{i_1 i_2} M_{i_2 i_3} \cdots M_{i_{k-1} i_k} M_{i_k i_1} , \tag{9}$$

and evaluating the leading contributions to the sum as $n_0 \to \infty$. Determining the leading contributions involves a complicated combinatorial analysis, combined with the evaluation of certain nontrivial high-dimensional integrals. In the next section and the supplementary material, we provide an outline for how to tackle these technical components of the computation.

3. The Stieltjes transform of M

3.1 Main result

The following theorem characterizes G as the solution to a quartic polynomial equation.

Theorem 1 *For M, ϕ, and ψ defined as in Section 2.1 and constants η and ζ defined as,*

$$\eta = \int dz \; \frac{e^{-z^2/2}}{\sqrt{2\pi}} f(\sigma_w \sigma_x z)^2 , \tag{10}$$

and,

$$\zeta = \left[\sigma_w \sigma_x \int dz \, \frac{e^{-z^2/2}}{\sqrt{2\pi}} f'(\sigma_w \sigma_x z)\right]^2 , \tag{11}$$

the Stieltjes transform of the spectral density of M satisfies,

$$G(z) = \frac{\psi}{z} P\left(\frac{1}{z\psi}\right) + \frac{1-\psi}{z} , \tag{12}$$

where,

$$P = 1 + (\eta - \zeta) t P_\phi P_\psi + \frac{P_\phi P_\psi t \zeta}{1 - P_\phi P_\psi t \zeta} , \tag{13}$$

and

$$P_\phi = 1 + (P-1)\phi , \quad P_\psi = 1 + (P-1)\psi . \tag{14}$$

The proof of Theorem 1 is relatively long and complicated, and it's deferred to the next section. The main idea in establishing the theorem is to translate the calculation of the moments in Equation 7 into two subproblems, one of enumerating certain connected outer-planar graphs, and another of evaluating integrals that correspond to cycles in those graphs. Much of the complexity resides in characterizing which outer-planar graphs contribute significantly to the moments and then computing those moments explicitly. A generating function encapsulating these results (P from Theorem 1) is shown to satisfy a relatively simple recurrence relation. Satisfying this recurrence relation requires that P solve Equation 13. Finally, some bookkeeping relates G to P. The details are presented in the supplementary material.

3.2 Limiting cases

3.2.1 $\eta = \zeta$

In Section 3 of the supplementary material, we use a Hermite polynomial expansion of f to show that $\eta = \zeta$ if and only if f is a linear function. In this case, $M = ZZ^T$, where $Z = WX$ is a product of Gaussian random matrices. Therefore we expect G to reduce to the Stieltjes transform of a so-called product Wishart matrix. In (Dupic and Castillo, 2014), a cubic equation defining the Stieltjes transform of such matrices is derived. Although Equation 12 is generally quartic, the coefficient of the quartic term vanishes when $\eta = \zeta$ (see Section 4 of the supplementary material). The resulting cubic polynomial is in agreement with the results in (Dupic and Castillo, 2014).

3.2.2 $\zeta = 0$

Another interesting limit is when $\zeta = 0$, which significantly simplifies the expression in Equation 13. Without loss of generality, we can take $\eta = 1$ (the general case can be recovered by rescaling z). The resulting equation is,

$$z\, G^2 + \Big((1 - \frac{\psi}{\phi})z - 1\Big) G + \frac{\psi}{\phi} = 0 , \tag{15}$$

which is precisely the Stieltjes transform of the Marchenko-Pastur distribution with shape parameter ϕ/ψ. Notice that when $\psi = 1$, the latter is the limiting spectral distribution

of XX^T, which implies that YY^T and XX^T have the same limiting spectral distribution. Therefore we have identified a novel type of isospectral non-linear transformation. We investigate this observation in Section 5.1.

4. Proof of Theorem 1

4.1 Polygonal Graphs

Theorem 1 relates the Stieltjes transform of G to the recurrence P, so our starting point is the equation of moments for G. Expanding out the powers of M in Equation 9 i.e., $\mathbf{E}\frac{1}{n_1}\operatorname{tr} M^k$, we have,

$$\mathbf{E}\,\frac{1}{n_1}\operatorname{tr} M^k = \frac{1}{n_1}\frac{1}{m^k}\,\mathbf{E}\sum_{\substack{i_1,\ldots,i_k\in[n_1]\\ \mu_1,\ldots,\mu_k\in[m]}} Y_{i_1\mu_1}Y_{i_2\mu_1}Y_{i_2\mu_2}Y_{i_3\mu_2}\cdots Y_{i_k\mu_k}Y_{i_1\mu_k}\,. \tag{16}$$

Notice that this sum can be decomposed based on the pattern of unique i and μ indices, and, because the elements of Y are i.i.d., the expected value of terms with the same index pattern is the same. Therefore, we are faced with the task of identifying the frequency of each index pattern and the corresponding expected values to leading order in n_0 as $n_0 \to \infty$.

To facilitate this analysis, it is useful to introduce a graph representation of the terms in Equation 16. For each term, i.e. each instantiation of indices i and μ in the sum, we will define a graph.

Consider first any term in which all indices are unique. In this case, we can identify each index with a vertex and each factor $Y_{i_j\mu_j}$ with an edge, and the corresponding graph can be visualized as a $2k$-sided polygon. There is a canonical planar embedding of such a cycle.

More generally, certain indices may be equal in the term. In this case, we can think of the term as corresponding to a polygonal cycle where certain vertices have been identified. The graph corresponding to such a term may be obtained by *pairwise identification* of the corresponding vertices in the canonical embedding of the $2k$-cycle. The graph now looks like a union of cycles, each joined to another at a common vertex.

Finally, admissible index identifications are those for which no i index is identified with a μ index and for which no pairings are crossing (with respect to the canonical embedding) The admissible graphs for $k = 3$ are shown in Figure 4.1, and for $k = 4$ in Figure 4.1.

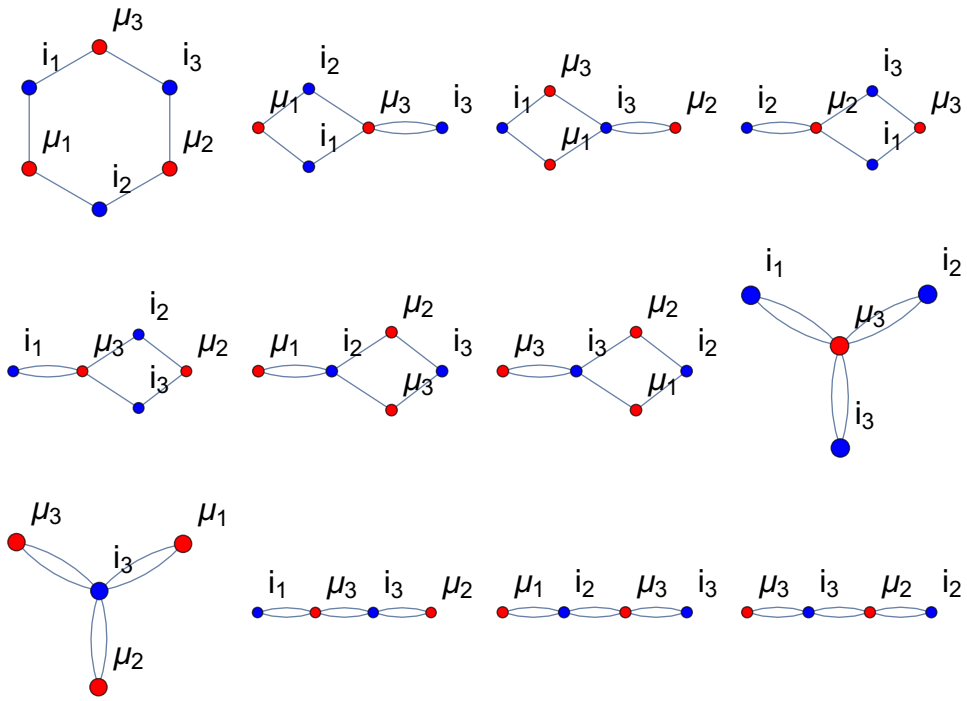


Figure 1: Admissible graphs for $k = 3$

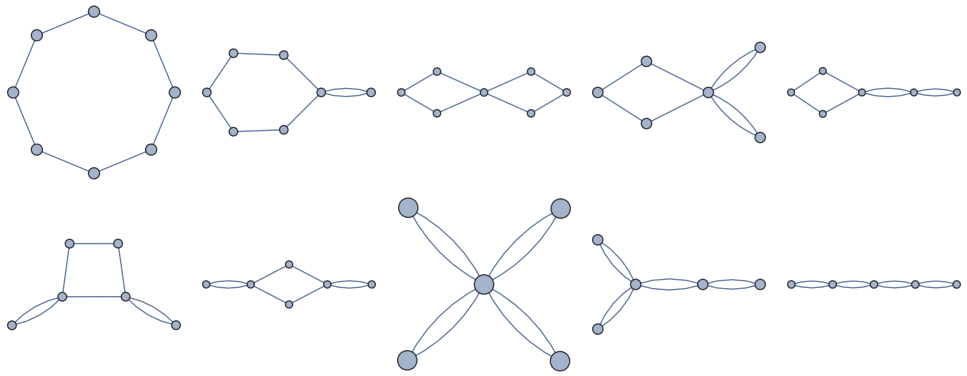


Figure 2: Admissible topologies for $k = 4$

To summarize the definitions more formally, we may define admissible graphs with $2k$ edges recursively.

Definition 2 *For any positive integer k, a $2k$-cycle, is an* admissible graph. *Start by labelling the vertices in the $2k$-cycle as $1, ..., 2k$ 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.*

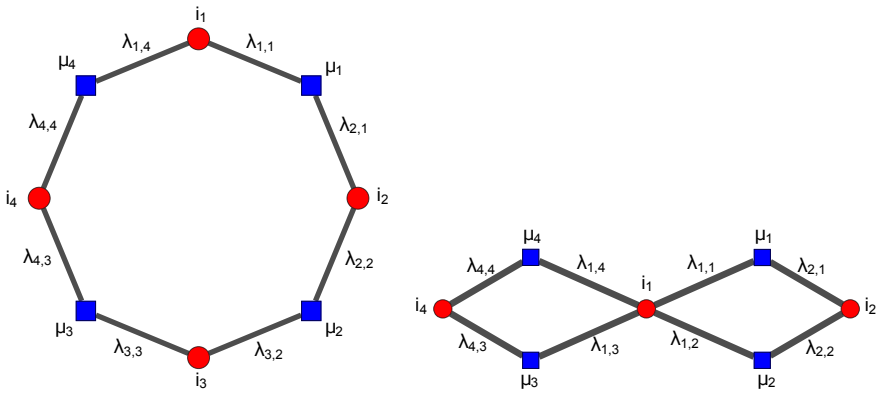


Figure 3: Two graphs for $2k = 8$.

Therefore, we have two ways to visualize admissible graphs. The first is only informally defined and can be thought of as certain kinds of "dissections" of a polygon. It serves as a tool for visualization. The second definition is algorithmic and defines the admissible graphs as an edge disjoint union of cycles. The latter view is useful for our proofs.

Proposition 3 *Every admissible graph is a connected outer-planar graph in which all blocks are simple even cycles.*

The proof follows from a simple inductive argument. In what follows, we will show that terms corresponding to such admissible graphs in Equation 16 determine the asymptotic value of the expectation.

4.2 Calculation of Moments

Let E_G denote the expected value of a term in Equation 16 corresponding to a graph G. We begin with the case where G is a $2k$-cycle. Each $2k$-cycle represents a multi-dimensional integral over the elements of W and X. Here we establish a correspondence between these integrals and a lower-dimensional integral whose structure is defined by the adjacency matrix of the graph. For a given $2k$-cycle, the expectation we wish to compute is,

$$\begin{aligned} E_{2k} &\equiv \mathbf{E}\, Y_{i_1\mu_1} Y_{i_2\mu_1} \cdots Y_{i_k\mu_k} Y_{i_1\mu_k} \qquad (17) \\ &= \int f(\sum_l W_{i_1 l} X_{l\mu_1}) f(\sum_l W_{i_2 l} X_{l\mu_1}) \cdots f(\sum_l W_{i_k l} X_{l\mu_k}) f(\sum_l W_{i_1 l} X_{l\mu_k})\, \mathcal{D}W\mathcal{D}X \end{aligned}$$

where,

$$\mathcal{D}W = \prod_{i,j=1}^{n_1,n_0} \frac{dW_{ij}}{\sqrt{2\pi\sigma_w^2/n_0}} e^{-\frac{n_0}{2\sigma_w^2} W_{ij}^2} \quad \mathcal{D}X = \prod_{i,\mu=1}^{n_0,d} \frac{dX_{i\mu}}{\sqrt{2\pi\sigma_x^2}} e^{-\frac{1}{2\sigma_x^2} X_{i\mu}^2}, \qquad (18)$$

and $i_1 \neq i_2 \neq ... \neq i_k \neq \mu_1 \neq \mu_2 \neq ... \neq \mu_k$. Next we introduce auxilliary integrals over z, which we can do by adding delta function contraints enforcing $Z = WX$. To this end, let $\mathcal{Z}$

denote the set of unique $Y_{i\mu}$ in Equation 17. Let $Z \in \mathbb{R}^{n_0 \times d}$ be the matrix whose entries are,

$$Z_{i\mu} = \begin{cases} z_{i\mu} & \text{if } Y_{i\mu} \in \mathcal{Z} \\ 0 & \text{otherwise}\,. \end{cases} \tag{19}$$

For each $y \in \mathcal{Z}$ we introduce an auxilliary integral,

$$E_{2k} = \int \prod_{z_{\alpha\beta} \in \mathcal{Z}} \delta(z_{\alpha\beta} - \sum_k W_{\alpha k} X_{k\beta})\; f(z_{i_1\mu_1}) f(z_{i_2\mu_1}) \cdots f(z_{i_k\mu_k}) f(z_{i_1\mu_k}) \mathcal{D}z \mathcal{D}W \mathcal{D}X\,, \tag{20}$$

where

$$\mathcal{D}z = \prod_{z_{\alpha\beta} \in \mathcal{Z}} dz_{\alpha\beta}\,. \tag{21}$$

Next we use a Fourier representation of the Dirac delta function,

$$\delta(x) = \frac{1}{2\pi} \int d\lambda\; e^{i\lambda x}\,, \tag{22}$$

for each of the delta functions in Equation 20. As above, we define a matrix $\Lambda \in \mathcal{R}^{n_1 \times d}$ whose entries are,

$$\Lambda_{i\mu} = \begin{cases} \lambda_{i\mu} & \text{if } Y_{i\mu} \in \mathcal{Z} \\ 0 & \text{otherwise}\,. \end{cases} \tag{23}$$

Then we can write,

$$E_{2k} = \int e^{-i\,\mathrm{tr}\,\Lambda^T(WX - Z)} f(z_{i_1\mu_1}) f(z_{i_2\mu_1}) \cdots f(z_{i_k\mu_k}) f(z_{i_1\mu_k})\; \mathcal{D}\lambda \mathcal{D}z \mathcal{D}W \mathcal{D}X, \tag{24}$$

where,

$$\mathcal{D}\lambda = \prod_{\lambda_{\alpha\beta} \in \Lambda} \frac{d\lambda_{\alpha\beta}}{2\pi}\,. \tag{25}$$

Note that the integral is bounded so we can use Fubini-Tonelli Theorem to switch integrals and perform the X and W integrals before λ and z integrals. We first perform the X integrals,

$$\begin{aligned} \int \mathcal{D}X\; e^{-i\,\mathrm{tr}\,\Lambda^T W X} &= \prod_{b,c=1}^{d,n_0} \int \frac{dX_{cb}}{\sqrt{2\pi\sigma_x^2}}\; \exp\left[-\frac{1}{2\sigma_x^2} X_{cb}^2 - i \sum_{a=1}^{n_1} \lambda_{ab} W_{ac} X_{cb} \right] \\ &= \exp\left[-\frac{\sigma_x^2}{2} \sum_{a,b,c=1}^{n_1,d,n_0} \left(\lambda_{ab} W_{ac}\right)^2 \right] \\ &= e^{-\frac{\sigma_x^2}{2}\,\mathrm{tr}\,\Lambda\Lambda^T W W^T}\,. \end{aligned} \tag{26}$$

Next we perform the W integrals,

$$\begin{aligned}\int \mathcal{D}W\; e^{-\frac{\sigma_x^2}{2}\operatorname{tr}\Lambda\Lambda^T WW^T} &= \prod_{i,j=1}^{n_1,n_0}\int \frac{dw_{i,j}}{(2\pi\sigma_w^2/n_0)^{1/2}}\; e^{-\frac{1}{2}\operatorname{tr}\left(\frac{n_0}{\sigma_w^2}I_{n_1}+\sigma_x^2\Lambda\Lambda^T\right)WW^T} \\ &= \prod_{j=1}^{n_0}\int \frac{d^{n_1}w_j}{(2\pi\sigma_w^2/n_0)^{n_1/2}}\; \exp\left[-\frac{1}{2}w_j^T\Big(\frac{n_0}{\sigma_w^2}I_{n_1}+\sigma_x^2\Lambda\Lambda^T\Big)w_j\right] \\ &= \prod_{i=1}^{n_0}\frac{1}{\det|I_{n_1}+\frac{\sigma_w^2\sigma_x^2}{n_0}\Lambda\Lambda^T|^{1/2}} \\ &= \frac{1}{\det|I_{n_1}+\frac{\sigma_w^2\sigma_x^2}{n_0}\Lambda\Lambda^T|^{n_0/2}}\,,\end{aligned} \tag{27}$$

where $w_j \in \mathbb{R}^{n_1}$ is the jth column of W and I_{n_1} is the $n_1 \times n_1$ identity matrix. Compiling the results up until now gives,

$$E_{2k} = \int \mathcal{D}\lambda\mathcal{D}z \frac{e^{-i\operatorname{tr}\Lambda Z}}{\det|1+\frac{\sigma_w^2\sigma_x^2}{n_0}\Lambda\Lambda^T|^{n_0/2}}F(z), \tag{28}$$

where we have introduced the abbreviation,

$$F(z) = f(z_{i_1\mu_1})f(z_{i_2\mu_1})\cdots f(z_{i_k\mu_k})f(z_{i_1\mu_k}) \tag{29}$$

to ease the notation. So far, we have not utilized the fact that n_0, n_1, and d are large. To proceed, we will use this fact to perform the λ integrals in the saddle point approximation, also known as the method of steepest descent. To this end, we write

$$E_{2k} = \int \mathcal{D}\lambda\mathcal{D}z\; \exp\left[-\frac{n_0}{2}\log\det|1+\frac{\sigma_w^2\sigma_x^2}{n_0}\Lambda\Lambda^T| - i\operatorname{tr}\Lambda Z\right]F(z) \tag{30}$$

and observe that the λ integrals will be dominated by contributions near where the coefficient of n_0 is minimized. It is straightforward to see that the minimizer is $\Lambda = 0$, at which point the phase factor $\operatorname{tr}\Lambda Z$ vanishes. Because the phase factor vanishes at the minimizer, we do not need to worry about the complexity of the integrand, and the approximation becomes equivalent to what is known as Laplace's method. The leading contributions to the integral come from the first non-vanishing terms in the expansion around the minimizer $\Lambda = 0$. To perform this expansion, we use the following identity, valid for small X[1],

$$\log\det|1+X| = \sum_{j=1}^{\infty}\frac{(-1)^{j+1}}{j}\operatorname{tr}X^j\,. \tag{31}$$

Using this expansion, we have,

$$\begin{aligned}E_{2k} &= \int \mathcal{D}\lambda\mathcal{D}z\; e^{-\frac{1}{2}\sigma_w^2\sigma_x^2\operatorname{tr}\Lambda\Lambda^T}e^{-\frac{n_0}{2}\sum_{j=2}^{\infty}\frac{(-1)^{j+1}}{j}\operatorname{tr}(\frac{\sigma_w^2\sigma_x^2}{n_0}\Lambda\Lambda^T)^j}e^{-i\operatorname{tr}\Lambda Z}F(z) \\ &= \int \mathcal{D}\tilde{\lambda}\mathcal{D}z\; e^{-\frac{n_0}{2}\operatorname{tr}\tilde{\Lambda}\tilde{\Lambda}^T}e^{-\frac{n_0}{2}\sum_{j=2}^{\infty}\frac{(-1)^{j+1}}{j}\operatorname{tr}(\tilde{\Lambda}\tilde{\Lambda}^T)^j}e^{-i\frac{\sqrt{n_0}}{\sigma_w\sigma_x}\operatorname{tr}\tilde{\Lambda}Z}F(z)\,,\end{aligned} \tag{32}$$

1. Since $\det|1+X|$ can be written as a polynomial $1+\operatorname{tr}(X)+O(\operatorname{tr}X^2)$, the expansion is valid as long as $\operatorname{tr}X < 1$. Plugging back Equation 30 for X, it suffices here that each entry in Λ goes to zero as $n_0 \to \infty$. Additionally, for Laplace's method we want the determinant of the Hessian of $\log\det|1+X|$ to deccay no faster exponential in n_0 Butler (2007). This is indeed the case here.

where we have changed integration variables to $\tilde{\lambda}_{ij} = \frac{\sigma_w \sigma_x}{\sqrt{n_0}} \lambda_{ij}$ and

$$\mathcal{D}\tilde{\lambda} = \prod_{\tilde{\lambda}_{\alpha\beta} \in \tilde{\Lambda}} \frac{d\tilde{\lambda}_{\alpha\beta}}{2\pi \sigma_w \sigma_x / \sqrt{n_0}} \,. \tag{33}$$

To extract the asymptotic contribution of this integral, we need to understand traces of $\tilde{\Lambda}\tilde{\Lambda}^T$. To this end, we make the following observation.

Lemma 4 *Given the matrix $\tilde{\Lambda} = [\tilde{\lambda}_{ij}]$, there exists matrix A such that*

$$\mathrm{tr}(\tilde{\Lambda}\tilde{\Lambda}^T)^k = \frac{1}{2} \,\mathrm{tr}\, A^{2k} \,, \tag{34}$$

where A is the weighted adjacency matrix defined by the undirected bigraph with vertex set $V = (I, U)$, where

$$I \equiv \{2i \,|\, \exists \mu \ s.t. \ y_{i\mu} \in \mathcal{Z}\}, \tag{35}$$

$$U \equiv \{2\mu - 1 \,|\, \exists i \ s.t. \ y_{i\mu} \in \mathcal{Z}\} \,, \tag{36}$$

and edges,

$$E \equiv \{\{2\mu - 1, 2i\} \,|\, y_{i\mu} \in \mathcal{Z}\} \,, \tag{37}$$

with weights $w(\{2\mu - 1, 2i\}) = \tilde{\lambda}_{i\mu}$.

The proof follows by defining an adjacency matrix:

$$A = \begin{pmatrix} 0 & \tilde{\Lambda} \\ \tilde{\Lambda}^T & 0 \end{pmatrix}, \tag{38}$$

where the I vertices are ordered before U vertices, and observe that the weights agree. Therefore,

$$A^{2k} = \begin{pmatrix} (\tilde{\Lambda}\tilde{\Lambda}^T)^k & 0 \\ 0 & (\tilde{\Lambda}^T\tilde{\Lambda})^k \end{pmatrix}. \tag{39}$$

Observe that the traces agree as required.

Now suppose that the middle exponential factor appearing in Equation 32 is truncated to finite order, m,

$$e^{-\frac{n_0}{2} \sum_{j=2}^{m} \frac{(-1)^{j+1}}{j} \mathrm{tr}(\tilde{\Lambda}\tilde{\Lambda}^T)^j} \,. \tag{40}$$

Since we are expanding for small $\tilde{\Lambda}$, we can expand the exponential into a polynomial of order $2m$. Any term in this polynomial that does not contain at least one factor $\tilde{\lambda}_{i\mu}$ for each $Y_{i\mu} \in \mathcal{Z}$ will vanish. To see this, denote (any one of) the missing $\tilde{\lambda}_{i\mu}$ as $\tilde{\lambda}$ and the corresponding $z_{i\mu}$ as z. Then,

$$\begin{aligned} \int dz \int \frac{d\tilde{\lambda}}{2\pi \sigma_w \sigma_x / \sqrt{n_0}} e^{-\frac{n_0}{2}\tilde{\lambda}^2} e^{-i\frac{\sqrt{n_0}}{\sigma_w \sigma_x}\tilde{\lambda} z} f(z) &= \int dz \, \frac{e^{-\frac{z^2}{2\sigma_w^2 \sigma_x^2}}}{\sqrt{2\pi \sigma_w^2 \sigma_x^2}} f(z) \\ &= \int \frac{dz}{\sqrt{2\pi}} \, e^{-\frac{z^2}{2}} f(\sigma_w \sigma_x z) \\ &= 0 \,, \end{aligned} \tag{41}$$

The last line follows from Equation 2.

The leading contribution to Equation 32 comes from the terms in the expansion of Equation 40 that have the fewest factors of $\tilde{\lambda}$, while still retaining one factor $\tilde{\lambda}_{i\mu}$ for each $Y_{i\mu}$. Since $\tilde{\lambda} \to 0$, as $n_0 \to \infty$ (the minimizer is $\tilde{\Lambda} = 0$), it follows that if there is a a term with exactly one factor of $\tilde{\lambda}_{i\mu}$ for each $Y_{i\mu}$, it will give the leading contribution. We now argue that there is always such a term, and we compute its coefficient.

Using Equation 34, traces of $\mathrm{tr}(\tilde{\Lambda}\tilde{\Lambda}^T)$ are equivalent to traces of A^2, where A is the adjacency matrix of the graph defined above. It is well known that the (u, v) entry of the A^k is the sum over weighted walks of length k, starting at vertex u and ending at vertex v. If there is a cycle of length k in the graph, then the diagonal elements of A^k contain two terms with exactly one factor of $\tilde{\lambda}$ for each edge in the cycle. (There are two terms arising from the clockwise and counter-clockwise walks around the cycle). Therefore, if there is a cycle of length $2k$, the expression $1/2 \,\mathrm{tr}\, A^{2k}$ contains a term with one factor of $\tilde{\lambda}$ for each edge in the cycle, with coefficient equal to $2k$.

So, finally, we can write for $k > 1$,

$$
\begin{aligned}
E_{2k} &= \int \mathcal{D}\tilde{\lambda}\mathcal{D}z \; e^{-\frac{n_0}{2}\,\mathrm{tr}\,\tilde{\Lambda}\tilde{\Lambda}^T} e^{-\frac{n_0}{2}\sum_{j=2}^{\infty}\frac{(-1)^{j+1}}{j}\,\mathrm{tr}(\tilde{\Lambda}\tilde{\Lambda}^T)^j} e^{-i\frac{\sqrt{n_0}}{\sigma_w\sigma_x}\,\mathrm{tr}\,\tilde{\Lambda}Z} F(z) \\
&\approx (-1)^k n_0 \int \mathcal{D}\tilde{\lambda}\mathcal{D}z \; e^{-\frac{n_0}{2}\,\mathrm{tr}\,\tilde{\Lambda}\tilde{\Lambda}^T} e^{-i\frac{\sqrt{n_0}}{\sigma_w\sigma_x}\,\mathrm{tr}\,\tilde{\Lambda}Z} \tilde{\lambda}_{i_1\mu_1}\tilde{\lambda}_{i_2\mu_1}\cdots\tilde{\lambda}_{i_k\mu_k}\tilde{\lambda}_{i_1\mu_k} F(z) \\
&= (-1)^k n_0 \left[\int \frac{d\tilde{\lambda}}{2\pi\sigma_w\sigma_x/\sqrt{n_0}} dz \; e^{-\frac{n_0}{2}\tilde{\lambda}^2} e^{-i\frac{\sqrt{n_0}}{\sigma_w\sigma_x}\tilde{\lambda}z}\tilde{\lambda}\, f(z)\right]^{2k} \\
&= (-1)^k n_0 \left[-i\int dz \; \frac{e^{-\frac{z^2}{2\sigma_w^2\sigma_x^2}}}{\sqrt{2\pi n_0}\sigma_w^2\sigma_x^2} z f(z)\right]^{2k} \\
&= n_0^{1-k}\left[\sigma_w\sigma_x\int dz \; \frac{e^{-z^2/2}}{\sqrt{2\pi}} f'(\sigma_w\sigma_x z)\right]^{2k} \\
&= n_0^{1-k}\zeta^k
\end{aligned}
\tag{42}
$$

where in the second to last line we have integrated by parts and we have defined,

$$
\zeta \equiv \left[\sigma_w\sigma_x\int dz \; \frac{e^{-z^2/2}}{\sqrt{2\pi}} f'(\sigma_w\sigma_x z)\right]^2 . \tag{43}
$$

We also note that if $k = 1$, there is no need to expand beyond first order because those integrals will not vanish (as they did in Equation 41). So in this case,

$$\begin{aligned} E_2 &\approx \int \mathcal{D}\tilde{\lambda}\mathcal{D}z \; e^{-\frac{n_0}{2}\,\mathrm{tr}\,\tilde{\Lambda}\tilde{\Lambda}^T} e^{-i\frac{\sqrt{n_0}}{\sigma_w\sigma_x}\,\mathrm{tr}\,\tilde{\Lambda}Z} F(z) \\ &= \left[\int \frac{d\tilde{\lambda}}{2\pi\sigma_w\sigma_x/\sqrt{n_0}} dz \; e^{-\frac{n_0}{2}\tilde{\lambda}^2} e^{-i\frac{\sqrt{n_0}}{\sigma_w\sigma_x}\tilde{\lambda}z} \, f(z)\right]^{2k} \\ &= \left[\int dz \; \frac{e^{-\frac{z^2}{2\sigma_w^2\sigma_x^2}}}{\sqrt{2\pi n_0\sigma_w^2\sigma_x^2}} f(z)\right]^{2k} \\ &= \int dz \; \frac{e^{-z^2/2}}{\sqrt{2\pi}} f(\sigma_w\sigma_x z)^2 \\ &\equiv \eta \,. \end{aligned} \tag{44}$$

The quantities η and ζ are important and will be used to obtain an expression for G.

The above was the simplest case, a $2k$ cycle. For any admissible graph G, we can view it as a tree over blocks, each block being a (even) cycle. If G has $2k$ edges then one can write the integral above as a product of integrals over edge disjoint cyclic blocks. More formally, suppose that G can be partitioned into edge disjoint (simple) even cycles $C_1, C_2, ..., C_c$ of length greater than 2 and b even cycles of length 2. The next step is to figure out the asymptotically dominant terms in the integral E_G in Equation 42. The difference in analysis, from the case when G is simple $2k$-cycle, will lie in the third equality. Each exponential term in the RHS of the second equality will be expressed as a product of $c + b$ disjoint sets of terms (in $\tilde{\lambda}$), each term corresponding to the edges of the c edge disjoint cyclic blocks and b "bubbles". Now we are back to the case of simple cycles of length 2 and above and we have shown the following statement.

Proposition 5 *Given an admissible graph G with c cyclic blocks, b blocks of size 1, and $2k$ edges, E_G grows as $n_0^{c-k} \cdot \eta^b \zeta^{c-b}$.*

4.2.1 Non-Admissible Graphs

In this subsection we consider the contribution of non-admissible graphs with $2k$-edges that are obtained by pairwise vertex identifications starting from the $2k$-cycle. Our aim in this subsection is to show that only the terms contributing to the admissible graphs determine the asymptotic value of the expectation in Equation 16. Observe that Equation 16 can have terms where subscript identifications of the form $i_j \leftrightarrow i_k$ and $\mu_j \leftrightarrow \mu_k$ can occur, but there are no identifications of the form $i_j \leftrightarrow \mu_k$ in our analysis. Therefore, in this subsection, it suffices to restrict our attention to terms corresponding to graphs that are non-admissible, but which obey the parity property (see Definition 2) i.e., the underlying graphs do not obey only the vertex connectivity property during vertex identifications.

Note that the number of terms (and therefore graphs) with k indices and c identifications is $\Theta(n_0^{2k-c})$. Although the fraction of non-admissible graphs with c identifications is far larger than that of admissible graphs as a function of k, the leading term for the integrals corresponding to latter grow as n_0^{c-k}, while the leading terms for the former grow at most

as n_0^{c-1-k}. The underlying reason for this sub-leading scaling is that any partition of a non-admissible graph into c blocks, where no two blocks have an edge in-between, requires c index identifications in the original $2k$-polygon, as opposed to $c-1$ identifications for an admissible graph. More formally, we have the following lemma and theorem.

Lemma 6 *A non-admissible graph with c vertex identifications can be partitioned into at most $c-1$ edge disjoint cyclic blocks.*

Proof Suppose the vertex connectivity property is violated after c' steps i.e., suppose that vertex v_1 and v_2 of the same partiy are identified in step c' but there exist only one vertex disjoint path between them. Now consider a partition of the resulting graph into edge disjoint cycles. In this case, the number of elements in such a partition remains c'-1. The reason being that the vertex connectivity between any two pairs of vertices in the graph does not decrease – a requirement to create a new element in the edge disjoint cycle partition. In general, each time the vertex connectivity property is violated during identifications, the number of edge disjoint blocks fails to increase. Therefore, any such non-admissible graph has a partition into at most $c-1$ edge disjoint blocks. ■

Theorem 7 *The value of E_G on a non-admissible graph with $2k$ vertices and c edge disjoint cycles is $O(n_0^{c-1-k})$.*

Proof Observe that the second integral in the LHS of Equation 42 can be expressed as a product of integrals over the elements in any edge disjoint cycle partition of the admissible graph. Each integral in the product can contribute a factor of n_0 and therefore the product integral is bounded above by $O(n_0^{c-1-k})$ (previous lemma). Therefore, even though, there are exponentially more non-admissible graphs as function of k, their net contribution in terms of n_0 is smaller. ■

Therefore, we may restrict our attention only to admissible graphs in order to complete the asymptotic evaluation of E_G.

4.3 Generating function

Let $\tilde{p}(k, v_i, v_\mu, b)$ denote the number of admissible graphs with $2k$ edges, v_i i-type vertex identifications, v_μ μ-type vertex identifications, and b cycles of size 1. Similarly, let $p(k, v_i, v_\mu, b)$ denote the same quantity modulo permutations of the vertices. Then, combining the definition of $G(z)$ (Equation 7) and Proposition 5, we have,

$$\begin{aligned} G(z) &\simeq \frac{1}{z} + \sum_{k=1}^{\infty} \sum_{v_i, v_\mu = 0}^{k} \sum_{b=0}^{v_i+v\mu+1} \binom{n_1}{k-v_i} \binom{m}{k-v_\mu} \frac{\tilde{p}(k, v_i, v_\mu, b)}{z^{k+1}} \frac{n_0^{v+v_\mu+1-k}}{n_1 m^k} \eta^b \zeta^{v_i+v\mu+1-b} \\ &\simeq \frac{1}{z} + \sum_{k=1}^{\infty} \frac{1}{z^{k+1}} \sum_{v_i, v_\mu = 0}^{k} \sum_{b=0}^{v_i+v_\mu+1} p(k, v_i, v_\mu, b) \eta^b \zeta^{v_i+v\mu+1-b} \phi^{v_i} \psi^{v_\mu} \\ &\simeq \frac{1-\psi}{z} + \frac{\psi}{z} \sum_{k=0}^{\infty} \frac{1}{z^k \psi^k} P(k), \end{aligned} \tag{45}$$

where we have defined,

$$P(k) = \sum_{v_i, v_\mu = 0}^{k} \sum_{b=0}^{v_i + v_\mu + 1} p(k, v_i, v_\mu, b) \eta^b \zeta^{v_i + v_\mu + 1 - b} \phi^{v_i} \psi^{v_\mu} \,. \tag{46}$$

Let $P(t) = \sum_k P(k) t^k$ be a generating function. Let $2k$ refer to the size of the cycle containing vertex 1. Summing over all possible values of k yields the following recurrence relation,

$$\begin{aligned} P &= 1 + t P_\phi P_\psi \eta + \sum_{k=2}^{\infty} (P_\phi P_\psi \zeta t)^k \\ &= 1 + (\eta - \zeta) t P_\phi P_\psi + \frac{P_\phi P_\psi t \zeta}{1 - P_\phi P_\psi t \zeta} \,. \end{aligned} \tag{47}$$

Note that if vertex 1 is inside a bubble, we get a factor of η instead of ζ, which is why that term is treated separately. The auxiliary generating functions P_ϕ and P_ψ correspond to the generating functions of graphs with an extra factor ϕ or ψ respectively, i.e.

$$P_\phi = 1 + (P - 1)\phi \quad P_\psi = 1 + (P - 1)\psi \,, \tag{48}$$

which arises from making a i-type or μ-type vertex identifications. Accounting for the relation between G and P in Equation 45 yields,

$$G(z) = \frac{\psi}{z} P\left(\frac{1}{z\psi}\right) + \frac{1 - \psi}{z} \,. \tag{49}$$

Hence, we have completed our proof of Theorem 1.

4.4 A combinatorial expression for moments

In the previous subsection, we almost obtained an explicit representation for $G(z)$ upto the undetermined function $p(k, v_i, v_\mu, b)$ in Equation 45 In this subsection we derive a combinatorial expression for the quantity p. While such an expression is not a requirement for the proof of Theorem 1, this subsection discusses the combinatorics of admissible graphs and that may be interesting its own right.

In this subsection, we will need to provide descriptions of counting arguments, and so to make the arguments easier to visualize, we will refer to i-type and μ-type vertex identifications as red and blue vertex identifications respectively.

Definition 8 *Given an admissible graph derived from a $2k$-cycle, let $\vec{\nu}(r, b, k)$ denote the* type vector *representing the number of cyclic blocks in the graph obtained after the r-red and b-blue vertex identifications. Therefore, ν_i gives the number of cyclic blocks of size $2i$ after the identifications, and it follows that:*

$$\sum_{i=1}^{k} 2i \cdot \nu_i = 2k, \tag{50}$$

$$\sum_{i=1}^{k} \nu_i = r + b + 1. \tag{51}$$

Definition 9 *Given an admissible graph derived from a $2k$-cycle with type vector $\vec{\nu}(r,b,k)$, define $\pi(\vec{\nu},r,b,k)$ to be the number of admissible graphs of type $\vec{\nu}$.*

Observe that $p(k,r,b,c)$ is simply a sum over type vectors as follows.

Proposition 10

$$p(k,r,b,c) = \sum_{\vec{\nu}:\nu_1=c} \pi(\vec{\nu}(r,b,k)). \tag{52}$$

An expression for p and G follows by taking further relevant summations.

Recall that admissible graphs can also be (roughly) thought of as vertex identifications within a polygon. Therefore, the problem of counting the number of admissible graphs of a given type $\vec{\nu}$ looks very similar to that of counting non-crossing polgonal dissections. Indeed the two problems are very related. Solutions to the latter typically proceed by construction of a bijection to planar trees, and such techniques, see for example Schuetz and Whieldon (2014) and the citations therein, can be used to count π as well, especially when the identifications are of one color only – one can set-up a bijection between admissible graphs and certain planar trees. However, with two or more colors the situation becomes more complicated. On the other hand, it is possible to count admissisble graphs of a given type by setting up a bijection directly with code-words and in what follows we will do just that.

The crux of the difference between vertex identifications for admissible graphs and more well-studied non-crossing dissections of polygons, as in Schuetz and Whieldon (2014), are addressed by the following example. Suppose, as in Figure 4.4, three vertices b, c and d are identified, which may be visualized as a non-crossing dissection as follows: We draw a dashed line between vertices b and c, and another dashed line between vertices c and d, one has a choice now – whether to draw a dashed line between b and d or not? To resolve the question, one may appeal to the algorithmic nature of the definition of admissible graphs. Recall that identifications are only made two at a time and such pairwise identification ensures that b and d will never be identified explicitly once b and c, and c and d are identified. Therefore, no line is drawn between b and d and the corresponding contribution to the type vector is at three coordinates only (as opposed to four).

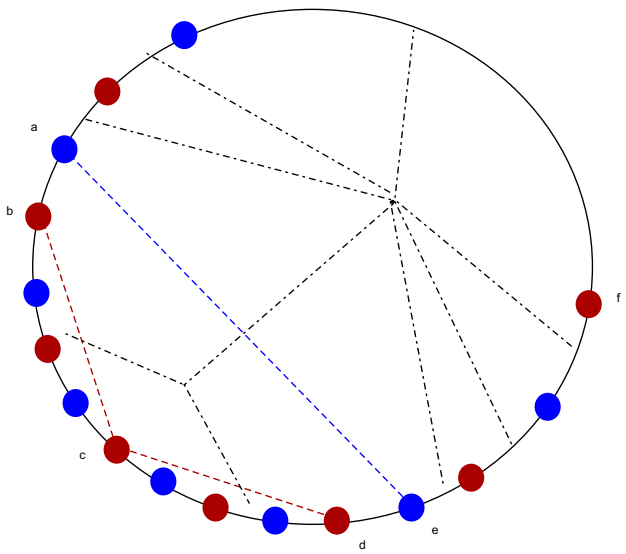


Figure 4: Admissible graph and its dual tree.

We first take a digression and state some combinatorial properties of admissble graphs which may help further understand their subtleties and then continue with deriving an expression for π.

4.4.1 Some combinatorial properties

Recall that admissible graphs are outer-planar. Moreover, if one considers a dual graph that is obtained by joining the vertex adjacent cyclic faces of an admissible graph[2] then one obtains a tree – the *dual tree*. For admissisble graphs, the leaves of such a dual tree correspond to the faces which share a vertex with only one other cyclic block. Furthermore, this dual tree has additional nice properties, some of which are listed below.

- Every face in the admissible graph i.e., every vertex of the tree, shares at least one edge with the outer face.

  Therefore, admissible graphs are not just outer-planar, but they are *"strongly" outer-planar*, in the above sense.

- Suppose one were to augment the dual tree mentioned above with extra vertices, such that each new vertex corresponds to an edge in the admissible graph. Next, add an edge in-between every new vertex and the vertex in the tree corresponding to the face

2. We exclude the outer face here.

containing the edge, then one obtains a planar embedding of a tree where every vertex has at least one adjacent leaf. In fact, every non-leaf face has an even number of leaves.

Therefore, for every admissible graph there corresponds one such *augmented dual tree*.

- Moreover, given a planar embedding of tree with $2k$ leaves, such that every non-leaf vertex has an even number of child as leaves, and that one extreme child of any non-leaf vertex is a leaf, then it is possible to recover the admissible graph. The idea is to simply replace every non-leaf vertex with an even sided polygon in a bottom-up fashion.

  Therefore, we have a bijection and our problem amounts to counting such a special class of planar trees as opposed to counting trees corresponding to non-crossing dissections in the usual sense. An expression for counting such trees may perhaps be interesting in itself.

An example: The augmented dual tree for the admissible graph in Figure 4.4 is illustrated by the dashed-dotted line. Face afe is the root (only the leaves on facet afe are drawn for convenience).

4.4.2 An expression for π

In this subsection, we will set up a bijection between admissible graphs of type $\vec{\nu}$ to the set of sequences $(k_1, s_1)(k_2, s_2)...(k_{r+b}, s_{r+b})$, where $k_1 \leq k_2 \leq k_3...$, and the multiplicity of any s_i in the sequence equals ν_{s_i}. It is easy to count such sequences, their number is $\binom{k}{r}\binom{k}{b}\binom{r+b}{\vec{\nu}}$. Here $\binom{r+b}{\vec{\nu}}$ stands for the multinomial coefficient $\binom{r+b}{\nu_1,...,\nu_k}$.

The bijection is best shown as an injection in both directions. The construction of the injections themselves is algorithmic in nature, which is why we prefer to work with admissible graphs as opposed to the corresponding class of trees.

Unfolding an admissible graph: Given an admissible graph we wish to construct a unique sequence $(k_1, s_1)(k_2, s_2)...(k_{r+b}, s_{r+b})$, as above.

We label the vertices in a counter-clockwise manner with $1, ..., 2k$, starting from an arbitrary fixed vertex. Recall that the pairwise identifications involved in generating an admissible graph are commutative, so we may assume that vertices are identified in increasing order of their indices. This gives a (unique) canonical labelling of the admissible graph.

We fix an arbitrary root face, say the one containing vertex 1, and associate with each cyclic block it's unique vertex which is closest to the root face. Suppose that v_i is the vertex for the cyclic face C_i and suppose that it was formed by the identification of vertices i_1 and i_2. We associate with v_i the minimum of i_1 and i_2 and denote it as k_i. Let s_i be the length of C_i. Therefore, we have associated a two-tuple (k_i, s_i) with C_i. Moreover, it is not hard to see that given an admissible graph and therefore it's unique canonical labelling this set of $r + b$ two-tuples is unique, which gives an injection in one direction.

Lemma 11

$$\pi(\vec{\nu}, r, b, k) \leq \frac{1}{r+b+1}\binom{k}{r}\binom{k}{b}\binom{r+b+1}{\vec{\nu}}. \tag{53}$$

Folding into an admissible graph: Our next task is to uniquely assign an endpoint on the $2k$ cycle to each element of the given a sequence $(k_1, s_1)(k_2, s_2)...(k_{r+b}, s_{r+b})$, as above, and therefore re-obtain the corresponding admissible graph.

The construction can be visualized as starting with a set of $r+b$ intervals on the line with points $1, ..., 2k$, each interval will eventually map (fold) into a cyclic face of the admissible graph. The start-point of interval i is the point k_i and s_i is the length of the cyclic face to be associated with the interval. To reconstruct the admissible graph, it suffices to find the end-point e_i for the interval (k_i, s_i), so that e_i is the (larger of two) indices for the vertex identification associated with a cyclic face C_i in the graph i.e., vertex indices k_i and e_i will be identified in the canonical labelling to obtain cyclic face C_i in the graph. Note that for the sequence to represent a canonically labelled admissible graph as above, it suffices that each length s_i is even and each of the start-points is unique and so is each of the end-points.

We can now find the end-point and thus reconstruct the canonically labelled admissible graph as follows:

Initially, let each $e_i := k_i + s_i$. For every pair of intersecting intervals, such that $k_i > k_j$ and $k_i < e_j$, set $e_j = e_j + s_i$ i.e., extend the end point of interval j by s_i. This may lead to new intersections, repeat the above procedure for any newly formed unaccounted intersections until no further new intersections are formed.

The end-result is a set of non intersecting even length intervals with distinct start and end-points in $1, ..., 2k$, as long as the total lengths of the initial set of intervals satisfies Equations 50 and 51. This gives a canonically labelled admissible graph. It is not too hard to see that the obtained admissible graph is unique and of type $\vec{\nu}$. Hence we have the following lemma.

Lemma 12

$$\pi(\vec{\nu}, r, b, k) \geq \frac{1}{r+b+1}\binom{k}{r}\binom{k}{b}\binom{r+b+1}{\vec{\nu}}. \tag{54}$$

Therefore, we have also shown the following theorem.

Theorem 13

$$\pi(\vec{\nu}, r, b, k) = \frac{1}{r+b+1}\binom{k}{r}\binom{k}{b}\binom{r+b+1}{\vec{\nu}}, \tag{55}$$

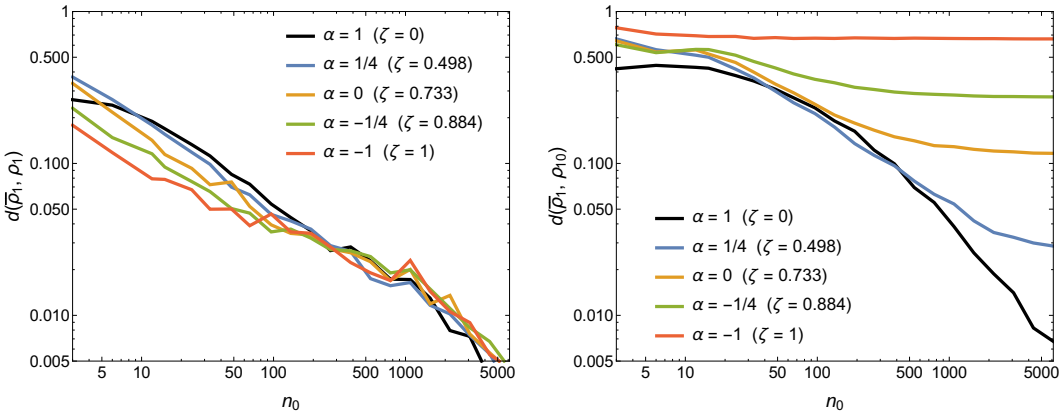
5. Applications

5.1 Data covariance

Consider a deep feed-forward neural network with lth layer post-activation matrix given by,

$$Y^l = f(W^l Y^{l-1}), \quad Y^0 = X\,. \tag{56}$$

The matrix $Y^l(Y^l)^T$ is the lth layer data covariance matrix. The distribution of its eigenvalues (or the singular values of Y^l) determine the extent to which the input signals become distorted or stretched as they propagate through the network. Highly skewed distributions indicate strong anisotropy in the embedded feature space, which is a form of poor conditioning that is likely to derail or impede learning. A variety of techniques have been developed to alleviate this problem, the most popular of which is batch normalization. In batch normalization, the variance of individual activations across the batch (or dataset) is rescaled to equal one. The covariance is often ignored — variants that attempt to fully whiten the activations can be very slow. So one aspect of batch normalization, as it is used in practice, is that it preserves



(a) $L = 1$ (b) $L = 10$

Figure 5: Distance between the (a) first-layer and (b) tenth-layer empirical eigenvalue distributions of the data covariance matrices and our theoretical prediction for the first-layer limiting distribution $\bar{\rho}_1$, as a function of network width n_0. Plots are for shape parameters $\phi = 1$ and $\psi = 3/2$. The different curves correspond to different piecewise linear activation functions parameterize by α: $\alpha = 1$ is linear, $\alpha = 0$ is (shifted) relu, and $\alpha = -1$ is (shifted) absolute value. In (a), for all α, we see good convergence of the empirical distribution ρ_1 to our asymptotic prediction $\bar{\rho}_1$. In (b), in accordance with our conjecture, we find good agreement between $\bar{\rho}_1$ and the tenth-layer empirical distribution $\zeta = 0$, but not for other values of ζ. This provides evidence that when $\zeta = 0$ the eigenvalue distribution is preserved by the non-linear transformations.

the trace of the covariance matrix (i.e. the first moment of its eigenvalue distribution) as the signal propagates through the network, but it does not control higher moments of the distribution. A consequence is that there may still be a large imbalance in singular values.

An interesting question, therefore, is whether there exist efficient techniques that could preserve or approximately preserve the full singular value spectrum of the activations as they propagate through the network. Inspired by the results of Section 3.2.2, we hypothesize that choosing an activation function with $\zeta = 0$ may be one way to approximately achieve this behavior, at least early in training. From a mathematical perspective, this hypothesis is similar to asking whether our results in Equation 12 are *universal* with respect to the distribution of X. We investigate this question empirically.

Let ρ_l be the empirical eigenvalue density of $Y^l(Y^l)^T$, and let $\bar{\rho}_1$ be the limiting density determined by Equation 12 (with $\psi = 1$). We would like to measure the distance between $\bar{\rho}_1$ and ρ_l in order to see whether the eigenvalues propagate without getting distorted. There are many options that would suffice, but we choose to track the following metric,

$$d(\bar{\rho}_1, \rho_l) \equiv \int d\lambda \, |\bar{\rho}_1(\lambda) - \rho_l(\lambda)| \; . \tag{57}$$

To observe the effect of varying ζ, we utilize a variant of the relu activation function with non-zero slope for negative inputs,

$$f_\alpha(x) = \frac{[x]_+ + \alpha[-x]_+ - \frac{1+\alpha}{\sqrt{2\pi}}}{\sqrt{\frac{1}{2}(1+\alpha^2) - \frac{1}{2\pi}(1+\alpha)^2}}\,. \tag{58}$$

One may interpret α as (the negative of) the ratio of the slope for negative x to the slope for positive x. It is straightforward to check that f_α has zero Gaussian mean and that,

$$\eta = 1, \quad \zeta = \frac{(1-\alpha)^2}{2(1+\alpha^2) - \frac{2}{\pi}(1+\alpha)^2}\,, \tag{59}$$

so we can adjust ζ (without affecting η) by changing α. Fig. 5(a) shows that for any value of α (and thus ζ) the distance between $\bar{\rho}_1$ and ρ_1 approaches zero as the network width increases. This offers numerical evidence that Equation 12 is in fact the correct asymptotic limit. It also shows how quickly the asymptotic behavior sets in, which is useful for interpreting Fig. 5(b), which shows the distance between $\bar{\rho}_1$ and ρ_{10}. Observe that if $\zeta = 0$, ρ_{10} approaches $\bar{\rho}_1$ as the network width increases. This provides evidence for the conjecture that the eigenvalues are in fact preserved as they propagate through the network, but only when $\zeta = 0$, since we see the distances level off at some finite value when $\zeta \neq 0$. We also note that small non-zero values of ζ may not distort the eigenvalues too much.

These observations suggest a new method of tuning the network for fast optimization. Choosing activation functions that satisfy $\zeta \approx 0$ could speed up training by equilibrating the singular values of the data covariance matrix as signals propagate through the network. We leave further investigation of these ideas to future work.

5.2 Asymptotic performance of random feature methods

Consider the ridge-regularized least squares loss function defined by,

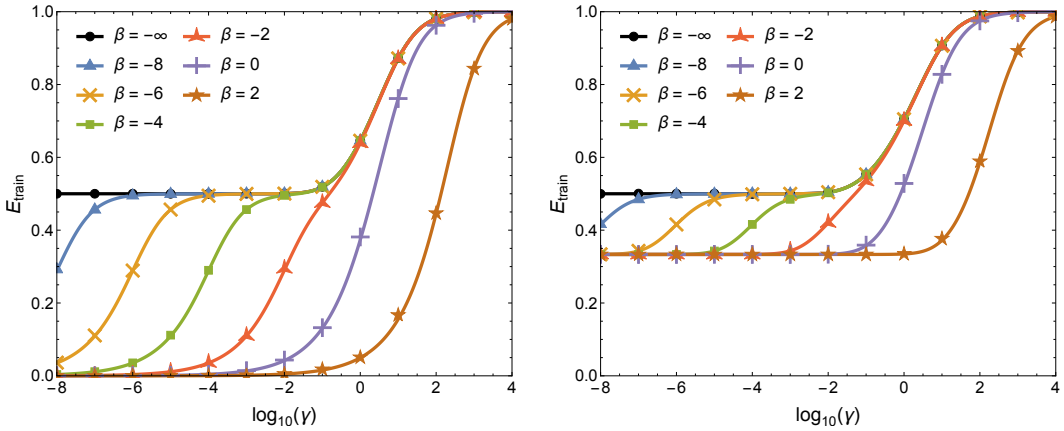
$$L(W_2) = \frac{1}{2n_2 m}\|\mathcal{Y} - W_2^T Y\|_F^2 + \gamma\|W_2\|_F^2, \quad Y = f(WX)\,, \tag{60}$$

where $X \in \mathbb{R}^{n_0 \times m}$ is a matrix of m n_0-dimensional features, $\mathcal{Y} \in \mathbb{R}^{n_2 \times m}$ is a matrix of regression targets, $W \in \mathbb{R}^{n_1 \times n_0}$ is a matrix of random weights and $W_2 \in \mathbb{R}^{n_1 \times n_2}$ is a matrix of parameters to be learned. The matrix Y is a matrix of random features[3]. The optimal parameters are,

$$W_2^* = \frac{1}{m} Y Q \mathcal{Y}^T, \quad Q = \left(\frac{1}{m} Y^T Y + \gamma I_m\right)^{-1}\,. \tag{61}$$

Our problem setup and analysis are similar to that of (Louart et al., 2017), but in contrast to that work, we are interested in the memorization setting in which the network is trained on random input-output pairs. Performance on this task is then a measure of the capacity of the model, or the complexity of the function class it belongs to. In this context, we take the

3. We emphasize that we are using an unconvential notation for the random features – we call them Y in order to make contact with the previous sections.



(a) $\phi = \frac{1}{2}$, $\psi = \frac{1}{2}$ (b) $\phi = \frac{1}{2}$, $\psi = \frac{3}{4}$

Figure 6: Memorization performance of random feature networks versus ridge regularization parameter γ. Theoretical curves in solid lines and numerical solutions to Equation 60 as points. $\beta \equiv \log_{10}(\eta/\zeta - 1)$ distinguishes classes of non-linearities, with $\beta = -\infty$ corresponding to a linear network. Each numerical simulation is done with a different function f with the specified β. The good agreement confirms that no details about f other than β are relevant. In (a), there are more random feature than data points, allowing for perfect memorization unless the function f is linear, in which case the model is rank constrained. In (b), there are fewer random features than data points, and even the non-linear models fail to achieve perfect memorization. The relative positions of the curves indicate that larger values of β (smaller values of ζ) improve the model's memorization capacity.

data X and the targets $\mathcal{Y}$ to be independent Gaussian random matrices. From eqns. (60) and (61), the expected training loss is given by,

$$\begin{aligned} E_{\text{train}} = \mathbf{E}_{W,X,\mathcal{Y}} L(W_2^*) &= \mathbf{E}_{W,X,\mathcal{Y}} \frac{\gamma^2}{m} \operatorname{tr} \mathcal{Y}^T \mathcal{Y} Q^2 \\ &= \mathbf{E}_{W,X} \frac{\gamma^2}{m} \operatorname{tr} Q^2 \\ &= -\frac{\gamma^2}{m} \frac{\partial}{\partial \gamma} \mathbf{E}_{W,X} \operatorname{tr} Q \,. \end{aligned} \tag{62}$$

It is evident from Equation 5 and the definition of Q that $\mathbf{E}_{W,X} \operatorname{tr} Q$ is related to $G(-\gamma)$. However, our results from the previous section cannot be used directly because Q contains the trace $Y^T Y$, whereas G was computed with respect to YY^T – not its transpose! Thankfully, the two matrices differ only by a finite number of zero eigenvalues. As in (Feinberg and Zee, 1997), some simple bookkeeping shows that

$$\frac{1}{m} \mathbf{E}_{W,X} \operatorname{tr} Q = \frac{(1 - \phi/\psi)}{\gamma} - \frac{\phi}{\psi} G(-\gamma) \,. \tag{63}$$

From Equation 12 and its total derivative with respect to z, an equation for $G'(z)$ can be obtained by computing the resultant of the two polynomials and eliminating $G(z)$. An equation for E_{train} follows; see the supplementary material for details. An analysis of this equation shows that it is homogeneous in γ, η, and ζ, i.e.,

$$E_{\text{train}}(\gamma, \eta, \zeta) = 0 \;\Rightarrow\; E_{\text{train}}(\lambda\gamma, \lambda\eta, \lambda\zeta) = 0\,. \tag{64}$$

Therefore it suffices to examine $E_{\text{train}}(\gamma, \eta, 1)$ since the general case can be recovered by choosing λ appropriately. The behavior when $\gamma = 0$ is a measure of the capacity of the model with no regularization and depends on the value of η,

$$E_{\text{train}}(0, \eta, 1) = \begin{cases} [1-\phi]_+ & \text{if } \eta = 1 \text{ and } \psi < 1, \\ [1-\phi/\psi]_+ & \text{otherwise.} \end{cases} \tag{65}$$

As discussed in Section 3.2, when $\eta = \zeta = 1$, the function f reduces to the identity. With this in mind, the various cases in Equation 65 are readily understood by considering the effective rank of the random feature matrix Y.

In fig. (6), we compare our theoretical predictions for E_{train} to numerical simulations of solutions to Equation 60. The different curves explore various ratios of $\beta \equiv \log_{10}(\eta/\zeta - 1)$ and therefore probe different classes of non-linearities. For each numerical simulation, we choose a random function f with the specified value of β (for details on this choice, see the supplementary material). The excellent agreement between theory and simulations confirms that E_{train} depends only on the parameters η (and ζ) and not on other details of f. The black curves correspond to the performance of a linear network. The results show that for η very close to 1, the models are unable to utilize their non-linearity unless the regularization parameter is very small. Conversely, for η very large, the models exploits the non-linearity very efficiently and can absorb large amounts of regularization without a significant drop in performance. This suggests that large η (or small ζ) might provide an interesting class of non-linear functions with enhanced expressive power.

6. Conclusions

In this work we studied the Gram matrix $M = \frac{1}{m} Y^T Y$, where $Y = f(WX)$ and W and X are random Gaussian matrices. We derived a quartic polynomial equation satisfied by the trace of the resolvent of M, which defines its limiting spectral density. In obtaining this result, we demonstrated that pointwise non-linearities can be incorporated into a standard method of proof in random matrix theory known as the moments method, thereby opening the door for future study of other non-linear random matrices appearing in neural networks.

We applied our results to a memorization task in the context of random feature methods and obtained an explicit characterizations of the training error as a function of a ridge regression parameter. The training error depends on the non-linearity only through two scalar quantities, η and ζ, which are certain Gaussian integrals of f. We observe that functions with small values of ζ appear to have increased capacity relative to those with larger values of ζ.

We also make the surprising observation that for $\zeta = 0$, the singular value distribution of $f(WX)$ is the same as the singular value distribution of X. In other words, the eigenvalues

of the data covariance matrix are constant in distribution when passing through a single non-linear layer of the network. We conjectured and found numerical evidence that this property actually holds when passing the signal through multiple layers. Therefore, we have identified a class of activation functions that maintains approximate isometry at initialization, which could have important practical consequences for training speed.

Both of our applications suggest that functions with $\zeta \approx 0$ are a potentially interesting class of activation functions. We note one crude way of interpreting ζ is as a measure of oddness of the function. Typically even activation functions are not used in deep learning, but maybe they should be.

Appendix

Appendix A. Hermite expansion

Any function with finite Gaussian moments can be expanded in a basis of permite polynomials. Defining

$$H_n(x) = (-1)^n e^{\frac{x^2}{2}} \frac{\partial^n}{\partial x^n} e^{-\frac{x^2}{2}} \tag{S1}$$

we can write,

$$f(x) = \sum_{n=0}^{\infty} \frac{f_n}{\sqrt{n!}} H_n(x)\,, \tag{S2}$$

for some constants f_n. Owing the orthogonality of the Hermite polynomials, this representation is useful for evaluating Gaussian integrals. In paticular, the condition that f be centered is equivalent the vanishing of f_0,

$$\begin{aligned} 0 &= \int_{-\infty}^{\infty} dx \frac{e^{-x^2/2}}{\sqrt{2\pi}} f(x) \\ &= f_0\,. \end{aligned} \tag{S3}$$

The constants η and ζ are also easily expressed in terms of the coefficients,

$$\begin{aligned} \eta &= \int_{-\infty}^{\infty} dx \frac{e^{-x^2/2}}{\sqrt{2\pi}} f(x)^2 \\ &= \sum_{n=0}^{\infty} f_n^2\,, \end{aligned} \tag{S4}$$

and,

$$\begin{aligned} \zeta &= \left[\int_{-\infty}^{\infty} dx \frac{e^{-x^2/2}}{\sqrt{2\pi}} f'(x) \right]^2 \\ &= f_1^2\,, \end{aligned} \tag{S5}$$

which together imply that $\eta \geq \zeta$. Equality holds when $f_{i>1} = 0$, in which case,

$$f(x) = f_1 H_1(x) = f_1 x \tag{S6}$$

i.e. when f is a linear function.

The Hermite representation also suggests a convenient way to randomly sample functions with specified values of η and ζ. First choose $f_1 = \sqrt{\zeta}$, and then enforce the constraint,

$$\eta - 1 = \sum_{n=2}^{N} f_n^2\,, \tag{S7}$$

where we have truncated the representation to some finite order N. Random values of f_n satisfying this relation are simple to obtain since they all live on the sphere of radius $\sqrt{\eta - 1}$.

Appendix B. Equations for Stieltjes transform

From Equation 12, straightforward algebra shows that G satisfies,

$$\sum_{i=0}^{4} a_i G^i = 0\,, \tag{S8}$$

where,

$$\begin{aligned}
a_0 &= -\psi^3\,,\\
a_1 &= \psi(\zeta(\psi-\phi)+\psi(\eta(\phi-\psi)+\psi z))\\
a_2 &= -\zeta^2(\phi-\psi)^2+\zeta\left(\eta(\phi-\psi)^2+\psi z(2\phi-\psi)\right)-\eta\psi^2 z\phi\\
a_3 &= \zeta(-z)\phi(2\zeta\psi-2\zeta\phi-2\eta\psi+2\eta\phi+\psi z)\\
a_4 &= \zeta z^2\phi^2(\eta-\zeta)\,.
\end{aligned} \tag{S9}$$

The total derivative of this equation with respect to z is,

$$\sum_{i=1}^{4} a_i' G^i + G'\sum_{i=0}^{3} b_i G^i = 0\,, \tag{S10}$$

where,

$$\begin{aligned}
a_1' &= \psi^3\,,\\
a_2' &= -\psi(\zeta(\psi-2\phi)+\eta\psi\phi)\,,\\
a_3' &= -2\zeta\phi(\zeta(\psi-\phi)+\eta(\phi-\psi)+\psi z)\,,\\
a_4' &= 2\zeta z\phi^2(\eta-\zeta)\,,\\
b_0 &= \psi(\zeta(\psi-\phi)+\psi(\eta(\phi-\psi)+\psi z))\\
b_1 &= 2\eta\left(\zeta(\phi-\psi)^2-\psi^2 z\phi\right)-2\zeta\left(\zeta(\phi-\psi)^2+\psi z(\psi-2\phi)\right)\\
b_2 &= -3\zeta z\phi(2\zeta\psi-2\zeta\phi-2\eta\psi+2\eta\phi+\psi z)\\
b_3 &= 4\zeta z^2\phi^2(\eta-\zeta)\,.
\end{aligned} \tag{S11}$$

To eliminate G from eqs. (S9) and (S11), we compute the resultant of the two polynomials, which produces a quartic polynomial in G'. Using eqns. (62) and (63) to change variables to E_{train}, we can derive the following equation satisfied by E_{train},

$$\sum_{i=0}^{4}\sum_{j=0}^{6} c_{i,j}\gamma^j E_{\text{train}}^i = 0\,, \tag{S12}$$

where the $c_{i,j}$ are given below. Notice that $\eta=\zeta$ is a degenerate case since $a_4 = b_3 = 0$ and the resultant must be computed separately. We find,

$$\sum_{i=0}^{3}\sum_{j=0}^{4} d_{i,j}\gamma^j E_{\text{train}}^i|_{\eta=\zeta} = 0\,, \tag{S13}$$

where the $d_{i,j}$ are given below. By inspection we find that

$$c_{i,j}(\lambda\eta, \lambda\zeta) = \lambda^{8-j} c_{i,j}(\eta, \zeta) \quad \text{and} \quad d_{i,j}(\lambda\zeta) = \lambda^{4-j} d_{i,j}(\zeta)\,, \tag{S14}$$

which establishes the homogeneity of E_{train} in γ, η, and ζ. From the coefficients $c_{i,0}$ we can read off the quartic equation satisfied by E_{train} when $\gamma = 0$ and $\eta \neq \zeta$. It has two double roots at,

$$E_{\text{train}}|_{\gamma=0} = 0 \quad \text{and} \quad E_{\text{train}}|_{\gamma=0} = 1 - \phi/\psi\,. \tag{S15}$$

In accordance with the condition that $G \to 1/z$ as $z \to \infty$, the first root is chosen if $\psi < \phi$ and the second root chosen otherwise.

If $\eta = \zeta$, then the coefficients $d_{i,0}$ define a cubic equation for E_{train} that has three distinct roots,

$$E_{\text{train}}|_{\gamma=0,\eta=\zeta} = 0\,, \quad E_{\text{train}}|_{\gamma=0,\eta=\zeta} = 1 - \phi\,, \quad \text{and} \quad E_{\text{train}}|_{\gamma=0,\eta=\zeta} = 1 - \phi/\psi\,. \tag{S16}$$

In this case, the first root is chosen when $\phi > \max(\psi, 1)$, the second root is chosen when $\phi, \psi < 1$, and the third root chosen otherwise.

Finally we give the coefficients $c_{i,j}$,

$$\begin{aligned} c_{0,0} = 0, \quad c_{0,1} = 0, \quad c_{0,2} = 0, \quad c_{0,3} = 0, \\ c_{1,0} = 0, \quad c_{1,1} = 0, \quad c_{3,6} = 0, \quad c_{4,6} = 0\,, \end{aligned}$$

and,

$$\begin{aligned} c_{0,4} &= \psi^6\phi^3\big(\zeta^2(4\psi - 1) - 2\zeta\eta\psi - \eta^2\psi^2\big)\big(\zeta^2\big((\psi - 1)\psi + \phi^2 + 2\psi\phi\big) - \\ &\qquad 2\zeta\eta\psi\phi - \eta^2\psi\phi^2\big) \\ c_{0,5} &= 2\zeta\psi^8\phi^3\big(\zeta^2(-\psi + \phi + 1) + \zeta\eta\big(-\psi^2 + \psi - 3\psi\phi + \phi\big) + \eta^2\psi\phi\big) \\ c_{0,6} &= -\zeta^2(\psi - 1)\psi^9\phi^3 \end{aligned}$$

$$
\begin{aligned}
c_{1,2} &= \psi^4\phi(\phi-\psi)^3(\zeta^2(4\psi-1)-2\zeta\eta\psi-\eta^2\psi^2)(\zeta^2(4\phi-1)-2\zeta\eta\phi-\\
&\quad \eta^2\phi^2)(\zeta^2(\psi+\phi-1)-\eta^2\psi\phi)\\
c_{1,3} &= -2\psi^5\phi(\phi-\psi)(\zeta^5(-(\psi-1)\psi^2-\phi^3+(-32\psi^2+9\psi+1)\phi^2+\psi(9\psi-4)\phi)+\\
&\quad \zeta^4\eta(-(\psi-1)\psi^3+(4\psi-1)\phi^4+(12\psi^2-8\psi+1)\phi^3+2\psi(6\psi^2+17\psi-2)\phi^2+4\psi^2(\psi^2-\\
&\quad 2\psi-1)\phi)-\zeta^3\eta^2\psi\phi((\psi-2)\psi^2+\phi^3+(7\psi-2)\phi^2+\psi(7\psi+8)\phi)+\\
&\quad 2\zeta^2\eta^3\psi\phi(\psi^3+(1-4\psi)\phi^3-4\psi^3\phi)+3\zeta\eta^4\psi^2\phi^2(\psi^2+\phi^2)+\\
&\quad \eta^5\psi^3\phi^3(\psi+\phi))\\
c_{1,4} &= \psi^6(-\phi)(\phi-\psi)(\zeta^4(-(\psi-1)\psi^2+(4\psi-1)\phi^3+(-16\psi^2+\psi+1)\phi^2+\psi(4\psi^2-\psi-\\
&\quad 9)\phi)+2\zeta^3\eta\psi\phi((\psi-1)\psi+\phi^2+(12\psi-1)\phi)+2\zeta^2\eta^2\psi\phi(3\psi^2+(3-9\psi)\phi^2+\\
&\quad (\psi-8\psi^2)\phi)+6\zeta\eta^3\psi^2\phi^2(\psi+\phi)+\eta^4\psi^3\phi^3)\\
c_{1,5} &= 2\zeta\psi^8\phi^2(\zeta^2((\psi-1)\psi-\phi^2+2\psi\phi+\phi)+2\zeta\eta(\psi^2+(2\psi-1)\phi^2-\psi^2\phi)+\\
&\quad \eta^2\psi\phi(\psi-\phi))\\
c_{1,6} &= \zeta^2\psi^9\phi^2(\psi+(\psi-1)\phi)
\end{aligned}
$$

$$
\begin{aligned}
c_{2,0} &= \zeta^2\psi^2(\zeta-\eta)^2(\phi-\psi)^6(\zeta^2(4\psi-1)-2\zeta\eta\psi-\eta^2\psi^2)(\zeta^2(4\phi-1)-\\
&\quad 2\zeta\eta\phi-\eta^2\phi^2)\\
c_{2,1} &= -2\zeta\psi^3(\zeta-\eta)(\phi-\psi)^4(\zeta^5(-5\psi^2+\psi+(16\psi-5)\phi^2+(16\psi^2-6\psi+1)\phi)+\\
&\quad \zeta^4\eta(-(\psi-3)\psi^2+(4\psi-1)\phi^3+(-40\psi^2-7\psi+3)\phi^2+\psi^2(4\psi-7)\phi)+\zeta^3\eta^2(2\psi^3+(2-\\
&\quad 9\psi)\phi^3+34\psi^2\phi^2-9\psi^3\phi)+\zeta^2\eta^3\psi\phi(\psi^2+(8\psi+1)\phi^2+2\psi(4\psi-3)\phi)-\\
&\quad 3\zeta\eta^4\psi^2\phi^2(\psi+\phi)-2\eta^5\psi^3\phi^3)\\
c_{2,2} &= \psi^4(-(\phi-\psi)^2)(\zeta^6(-\psi^2(\psi^2-8\psi+1)+(4\psi-1)\phi^4+(-148\psi^2+22\psi+8)\phi^3-\\
&\quad (148\psi^3-118\psi^2+21\psi+1)\phi^2+\psi(4\psi^3+22\psi^2-21\psi+3)\phi)-2\zeta^5\eta(-3(\psi-1)\psi^3+(11\psi-\\
&\quad 3)\phi^4+(-147\psi^2+24\psi+3)\phi^3+\psi(-147\psi^2+66\psi-8)\phi^2+\psi^2(11\psi^2+24\psi-8)\phi)+\\
&\quad \zeta^4\eta^2(-6\psi^4+(66\psi^2+9\psi-6)\phi^4+\psi(28\psi^2-199\psi+27)\phi^3+\psi^2(66\psi^2-199\psi+29)\phi^2+\\
&\quad 9\psi^3(\psi+3)\phi)+2\zeta^3\eta^3\psi\phi(5\psi^3+(5-44\psi)\phi^3+(21-20\psi)\psi\phi^2+(21-44\psi)\psi^2\phi)+\\
&\quad \zeta^2\eta^4\psi^2\phi^2(24\psi^2+(24-13\psi)\phi^2+(23-13\psi)\psi\phi)+10\zeta\eta^5\psi^3\phi^3(\psi+\phi)+\\
&\quad \eta^6\psi^4\phi^4)
\end{aligned}
$$

$$
\begin{aligned}
c_{2,3} = {} & 2\psi^5(\zeta^5(-(\psi-1)\psi^4+(3\psi-1)\phi^5+(-36\psi^2+19\psi+1)\phi^4+\psi(98\psi^2-26\psi-7)\phi^3- \\
& 2\psi^2(18\psi^2+13\psi-7)\phi^2+\psi^3(3\psi^2+19\psi-7)\phi)+\zeta^4\eta(2\psi^5+(-40\psi^2+5\psi+2)\phi^5+ \\
& \psi(24\psi^2+54\psi-19)\phi^4+2\psi^2(12\psi^2-67\psi+10)\phi^3+2\psi^3(-20\psi^2+27\psi+10)\phi^2+\psi^4(5\psi- \\
& 19)\phi)+\zeta^3\eta^2\psi\phi(-11\psi^4+(50\psi-11)\phi^4-2\psi(21\psi+1)\phi^3-6\psi^2(7\psi-5)\phi^2+2\psi^3(25\psi- \\
& 1)\phi)+2\zeta^2\eta^3\psi^2\phi^2(-7\psi^3+(5\psi-7)\phi^3-2(\psi-3)\psi\phi^2+\psi^2(5\psi+6)\phi)+ \\
& \zeta\eta^4\psi^3\phi^3(-5\psi^2-5\phi^2+4\psi\phi)-\eta^5\psi^4\phi^4(\psi+\phi)) \\
c_{2,4} = {} & \psi^6(\zeta^4(\psi^4+(-31\psi^2+7\psi+1)\phi^4+\psi(70\psi^2-6\psi-13)\phi^3+\psi^2(-31\psi^2-6\psi+ \\
& 31)\phi^2+\psi^3(7\psi-13)\phi)+2\zeta^3\eta\psi\phi(-8\psi^3+(17\psi-8)\phi^3+3(3-16\psi)\psi\phi^2+\psi^2(17\psi+ \\
& 9)\phi)+\zeta^2\eta^2\psi^2\phi^2(-14\psi^2+(17\psi-14)\phi^2+\psi(17\psi+14)\phi)-6\zeta\eta^3\psi^3\phi^3(\psi+\phi)- \\
& \eta^4\psi^4\phi^4) \\
c_{2,5} = {} & -2\zeta\psi^8\phi(\zeta^2(2\psi^2+(\psi+2)\phi^2+(\psi-5)\psi\phi)+\zeta\eta\psi\phi(2\psi+(2-3\psi)\phi)+ \\
& \eta^2\psi^2\phi^2) \\
c_{2,6} = {} & -\zeta^2\psi^{10}\phi^2
\end{aligned}
$$

$$
\begin{aligned}
c_{3,0} = {} & 2\zeta^2\psi^3(\zeta-\eta)^2(\phi-\psi)^5(\zeta^2(4\psi-1)-2\zeta\eta\psi-\eta^2\psi^2)(\zeta^2(4\phi-1)- \\
& 2\zeta\eta\phi-\eta^2\phi^2) \\
c_{3,1} = {} & -4\zeta\psi^4(\zeta-\eta)(\phi-\psi)^3(\zeta^5(-5\psi^2+\psi+(16\psi-5)\phi^2+(16\psi^2-6\psi+1)\phi)+ \\
& \zeta^4\eta(-(\psi-3)\psi^2+(4\psi-1)\phi^3+(-40\psi^2-7\psi+3)\phi^2+\psi^2(4\psi-7)\phi)+\zeta^3\eta^2(2\psi^3+(2- \\
& 9\psi)\phi^3+34\psi^2\phi^2-9\psi^3\phi)+\zeta^2\eta^3\psi\phi(\psi^2+(8\psi+1)\phi^2+2\psi(4\psi-3)\phi)- \\
& 3\zeta\eta^4\psi^2\phi^2(\psi+\phi)-2\eta^5\psi^3\phi^3) \\
c_{3,2} = {} & -2\zeta\psi^5(\phi-\psi)(\zeta^5(-\psi^2(\psi^2-8\psi+1)+(4\psi-1)\phi^4+(-132\psi^2+18\psi+8)\phi^3- \\
& (132\psi^3-94\psi^2+16\psi+1)\phi^2+2\psi(2\psi^3+9\psi^2-8\psi+1)\phi)-2\zeta^4\eta(-3(\psi-1)\psi^3+(11\psi- \\
& 3)\phi^4+(-139\psi^2+23\psi+3)\phi^3+\psi(-139\psi^2+56\psi-7)\phi^2+\psi^2(11\psi^2+23\psi-7)\phi)+ \\
& 2\zeta^3\eta^2(-3\psi^4+(31\psi^2+5\psi-3)\phi^4+\psi(2\psi^2-93\psi+13)\phi^3+\psi^2(31\psi^2-93\psi+12)\phi^2+ \\
& \psi^3(5\psi+13)\phi)+2\zeta^2\eta^3\psi\phi(5\psi^3+(5-43\psi)\phi^3+(19-10\psi)\psi\phi^2+(19-43\psi)\psi^2\phi)+ \\
& \zeta\eta^4\psi^2\phi^2(23\psi^2+(23-8\psi)\phi^2+2(9-4\psi)\psi\phi)+8\eta^5\psi^3\phi^3(\psi+\phi))
\end{aligned}
$$

$$
\begin{aligned}
c_{3,3} &= 4\zeta\psi^6(\phi-\psi)(\zeta^4(-(\psi-1)\psi^2+(3\psi-1)\phi^3+(-30\psi^2+16\psi+1)\phi^2+\psi(3\psi^2+16\psi-\\
&\quad 4)\phi)+2\zeta^3\eta(\psi^3+(-18\psi^2+2\psi+1)\phi^3+\psi(-18\psi^2+27\psi-7)\phi^2+\psi^2(2\psi-7)\phi)+\\
&\quad \zeta^2\eta^2\psi\phi(-11\psi^2+(49\psi-11)\phi^2+\psi(49\psi-22)\phi)+2\zeta\eta^3\psi^2\phi^2((\psi-6)\phi-6\psi)-\\
&\quad 2\eta^4\psi^3\phi^3)\\
c_{3,4} &= -2\zeta^2\psi^7(\phi-\psi)(\zeta^2(-\psi^2+(27\psi^2-6\psi-1)\phi^2+2(5-3\psi)\psi\phi)-\\
&\quad 4\zeta\eta\psi\phi((9\psi-4)\phi-4\psi)+8\eta^2\psi^2\phi^2)\\
c_{3,5} &= 8\zeta^3\psi^9\phi(\psi-\phi)
\end{aligned}
$$

$$
\begin{aligned}
c_{4,0} &= \zeta^2\psi^4(\zeta-\eta)^2(\phi-\psi)^4(\zeta^2(4\psi-1)-2\zeta\eta\psi-\eta^2\psi^2)(\zeta^2(4\phi-1)-\\
&\quad 2\zeta\eta\phi-\eta^2\phi^2)\\
c_{4,1} &= -2\zeta\psi^5(\zeta-\eta)(\phi-\psi)^2(\zeta^5(-5\psi^2+\psi+(16\psi-5)\phi^2+(16\psi^2-6\psi+1)\phi)+\\
&\quad \zeta^4\eta(-(\psi-3)\psi^2+(4\psi-1)\phi^3+(-40\psi^2-7\psi+3)\phi^2+\psi^2(4\psi-7)\phi)+\zeta^3\eta^2(2\psi^3+(2-\\
&\quad 9\psi)\phi^3+34\psi^2\phi^2-9\psi^3\phi)+\zeta^2\eta^3\psi\phi(\psi^2+(8\psi+1)\phi^2+2\psi(4\psi-3)\phi)-\\
&\quad 3\zeta\eta^4\psi^2\phi^2(\psi+\phi)-2\eta^5\psi^3\phi^3)\\
c_{4,2} &= -\zeta\psi^6(\zeta^5(-\psi^2(\psi^2-8\psi+1)+(4\psi-1)\phi^4+(-132\psi^2+18\psi+8)\phi^3-(132\psi^3-94\psi^2\\
&\quad +16\psi+1)\phi^2+2\psi(2\psi^3+9\psi^2-8\psi+1)\phi)-2\zeta^4\eta(-3(\psi-1)\psi^3+(11\psi-3)\phi^4+(-139\psi^2+\\
&\quad 23\psi+3)\phi^3+\psi(-139\psi^2+56\psi-7)\phi^2+\psi^2(11\psi^2+23\psi-7)\phi)+2\zeta^3\eta^2(-3\psi^4+\\
&\quad (31\psi^2+5\psi-3)\phi^4+\psi(2\psi^2-93\psi+13)\phi^3+\psi^2(31\psi^2-93\psi+12)\phi^2+\psi^3(5\psi+13)\phi)+\\
&\quad 2\zeta^2\eta^3\psi\phi(5\psi^3+(5-43\psi)\phi^3+(19-10\psi)\psi\phi^2+(19-43\psi)\psi^2\phi)+\\
&\quad \zeta\eta^4\psi^2\phi^2(23\psi^2+(23-8\psi)\phi^2+2(9-4\psi)\psi\phi)+8\eta^5\psi^3\phi^3(\psi+\phi))\\
c_{4,3} &= 2\zeta\psi^7(\zeta^4(-(\psi-1)\psi^2+(3\psi-1)\phi^3+(-30\psi^2+16\psi+1)\phi^2+\psi(3\psi^2+16\psi-\\
&\quad 4)\phi)+2\zeta^3\eta(\psi^3+(-18\psi^2+2\psi+1)\phi^3+\psi(-18\psi^2+27\psi-7)\phi^2+\psi^2(2\psi-7)\phi)+\\
&\quad \zeta^2\eta^2\psi\phi(-11\psi^2+(49\psi-11)\phi^2+\psi(49\psi-22)\phi)+2\zeta\eta^3\psi^2\phi^2((\psi-6)\phi-6\psi)-\\
&\quad 2\eta^4\psi^3\phi^3)\\
c_{4,4} &= \zeta^2\psi^8(\zeta^2(\psi^2+(-27\psi^2+6\psi+1)\phi^2+2\psi(3\psi-5)\phi)+4\zeta\eta\psi\phi((9\psi-4)\phi-\\
&\quad 4\psi)-8\eta^2\psi^2\phi^2)\\
c_{4,5} &= -4\zeta^3\psi^{10}\phi
\end{aligned}
$$

And the coefficients $d_{i,j}$ read,

$$d_{0,0}=0,\quad d_{0,1}=0,\quad d_{2,4}=0,\quad d_{3,4}=0\,,$$

and,

$$
\begin{aligned}
d_{0,2} &= -\zeta^2(\psi-1)^2\psi^2\phi^2(\phi^2-\psi)\\
d_{0,3} &= 2\zeta\psi^4\phi^2(\psi+2\phi+1)\\
d_{0,4} &= \psi^5\phi^2
\end{aligned}
$$

$$
\begin{aligned}
d_{1,0} &= \zeta^4(\psi-1)^2(\phi-1)^3(\phi-\psi)^3\\
d_{1,1} &= 2\zeta^3\psi(\phi-1)\big(-\psi^3(\psi+1)+(\psi^2-4\psi+1)\phi^4+(6\psi^2+\psi+1)\phi^3-\psi(\psi^3+6\psi^2+\\
&\qquad 5)\phi^2+\psi^2(4\psi^2-\psi+5)\phi\big)\\
d_{1,2} &= \zeta^2\psi^2\big(\psi^3+(\psi^2-11\psi+1)\phi^4-(\psi^3+1)\phi^3+2\psi(5\psi^2+6\psi+5)\phi^2-\\
&\qquad 11\psi^2(\psi+1)\phi\big)\\
d_{1,3} &= 2\zeta\psi^4\phi\big(-2\psi-3\phi^2+\psi\phi+\phi\big)\\
d_{1,4} &= \psi^5\big(-\phi^2\big)
\end{aligned}
$$

$$
\begin{aligned}
d_{2,0} &= \zeta^4(\psi-1)^2(\phi-1)^2(\phi-\psi)^2(-2\psi+\psi\phi+\phi)\\
d_{2,1} &= 2\zeta^3\psi\big(-2\psi^3(\psi+1)+(\psi^3-3\psi^2-3\psi+1)\phi^4+(\psi^4+\psi^3+12\psi^2+\psi+1)\phi^3-\\
&\qquad 6\psi(\psi^3+\psi^2+\psi+1)\phi^2+\psi^2(9\psi^2-2\psi+9)\phi\big)\\
d_{2,2} &= \zeta^2\psi^2\big(-2\psi^3+(\psi^3-9\psi^2-9\psi+1)\phi^3-12(\psi^3+\psi)\phi^2+21\psi^2(\psi+1)\phi\big)\\
d_{2,3} &= -4\zeta\psi^4\phi(-2\psi+\psi\phi+\phi)
\end{aligned}
$$

$$
\begin{aligned}
d_{3,0} &= \zeta^4(\psi-1)^2\psi(\phi-1)^2(\phi-\psi)^2\\
d_{3,1} &= 2\zeta^3\psi^2\big(\psi^2(\psi+1)+(\psi^2-4\psi+1)\phi^3+(\psi^3+2\psi^2+2\psi+1)\phi^2+2\psi\big(-2\psi^2+\psi-\\
&\qquad 2)\phi\big)\\
d_{3,2} &= \zeta^2\psi^3\big(\psi^2+(\psi^2-10\psi+1)\phi^2-10\psi(\psi+1)\phi\big)\\
d_{3,3} &= -4\zeta\psi^5\phi
\end{aligned}
$$

References

Han Altae-Tran, Bharath Ramsundar, Aneesh S. Pappu, and Vijay Pande. Low Data Drug Discovery with One-Shot Learning. *American Chemical Society Central Science*, 2017.

Alan Bray and David Dean. tatistics of critical points of gaussian fields on large-dimensional spaces. *Physical Review Letters*, 2007.

Ronald Butler. *Saddlepoint approximations and applications*. Cambridge University Press, 2007.

Xiuyuan Cheng and Amit Singer. The spectrum of random inner-product kernel matrices. *Random Matrices: Theory and Applications*, 2(04):1350010, 2013.

Anna Choromanska, Mikael Henaff, Michael Mathieu, Gérard Ben Arous, and Yann LeCun. The loss surfaces of multilayer networks. In *AISTATS*, 2015.

A. Daniely, R. Frostig, and Y. Singer. Toward Deeper Understanding of Neural Networks: The Power of Initialization and a Dual View on Expressivity. *arXiv:1602.05897*, 2016.

Yann Dauphin, Razvan Pascanu, Çaglar Gülçehre, Kyunghyun Cho, Surya Ganguli, and Yoshua Bengio. Identifying and attacking the saddle point problem in high-dimensional non-convex optimization. *CoRR*, abs/1406.2572, 2014. URL `http://arxiv.org/abs/1406.2572`.

Thomas Dupic and Isaac Pérez Castillo. Spectral density of products of wishart dilute random matrices. part i: the dense case. *arXiv preprint arXiv:1401.7802*, 2014.

Noureddine El Karoui et al. The spectrum of kernel random matrices. *The Annals of Statistics*, 38(1):1–50, 2010.

Joshua Feinberg and Anthony Zee. Renormalizing rectangles and other topics in random matrix theory. *Journal of statistical physics*, 87(3-4):473–504, 1997.

Yan Fyodorov and Ian Williams. Replica symmetry breaking condition exposed by random matrix calculation of landscape complexity. *Journal of Statistical Physics*, 2007.

Garrett Goh, Nathan Hodas, and Abhinav Vishnu. Deep Learning for Computational Chemistry. *arXiv preprint arXiv:1701.04503*, 2017.

Geoffrey Hinton, Li Deng, Dong Yu, George E. Dahl, Abdel-rahman Mohamed, Navdeep Jaitly, Andrew Senior, Vincent Vanhoucke, Patrick Nguyen, Tara N Sainath, et al. Deep neural networks for acoustic modeling in speech recognition: The shared views of four research groups. *IEEE Signal Processing Magazine*, 29(6):82–97, 2012.

Sergey Ioffe and Christian Szegedy. Batch normalization: Accelerating deep network training by reducing internal covariate shift. In *Proceedings of The 32nd International Conference on Machine Learning*, pages 448–456, 2015.

Alex Krizhevsky, Ilya Sutskever, and Geoffrey E Hinton. Imagenet classification with deep convolutional neural networks. In *Advances in neural information processing systems*, pages 1097–1105, 2012.

Cosme Louart, Zhenyu Liao, and Romain Couillet. A random matrix approach to neural networks. *arXiv preprint arXiv:1702.05419*, 2017.

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.

Radford M Neal. *Bayesian learning for neural networks*, volume 118. Springer Science & Business Media, 2012.

Aaron van den Oord, Sander Dieleman, Heiga Zen, Karen Simonyan, Oriol Vinyals, Alex Graves, Nal Kalchbrenner, Andrew Senior, and Koray Kavukcuoglu. Wavenet: A generative model for raw audio. *arXiv preprint arXiv:1609.03499*, 2016.

Baldi Pierre and Homik Kurt. Neural networks and principal component analysis: Learning from examples without local minima. *Neural Networks*, 1989.

B. Poole, S. Lahiri, M. Raghu, J. Sohl-Dickstein, and S. Ganguli. Exponential expressivity in deep neural networks through transient chaos. *arXiv:1606.05340*, June 2016.

M. Raghu, B. Poole, J. Kleinberg, S. Ganguli, and J. Sohl-Dickstein. On the expressive power of deep neural networks. *arXiv:1606.05336*, June 2016.

Ali Rahimi and Ben Recht. Random features for large-scale kernel machines. In *In Neural Infomration Processing Systems*, 2007.

David Saad. *Online learning in neural networks*. Cambridge University Press, 2009.

David Saad and Sara Solla. Exact solution for online learning in multilayer neural networks. *Physical Review Letters*, 1995.

A. M. Saxe, J. L. McClelland, and S. Ganguli. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *International Conference on Learning Representations*, 2014.

S. S. Schoenholz, J. Gilmer, S. Ganguli, and J. Sohl-Dickstein. Deep Information Propagation. *ArXiv e-prints*, November 2016.

Alison Schuetz and Gwyn Whieldon. Polgonal dissections and reversions of series. *arXiv preprint arXiv:1401.7194*, 2014.

N. Shazeer, A. Mirhoseini, K. Maziarz, A. Davis, Q. Le, G. Hinton, and J. Dean. Outrageously large neural language models using sparsely gated mixtures of experts. *ICLR*, 2017. URL `http://arxiv.org/abs/1701.06538`.

Terence Tao. *Topics in random matrix theory*, volume 132. American Mathematical Society Providence, RI, 2012.

Yonghui Wu, Mike Schuster, Zhifeng Chen, Quoc V. Le, Mohammad Norouzi, Wolfgang Macherey, Maxim Krikun, Yuan Cao, Qin Gao, Klaus Macherey, et al. Google's neural machine translation system: Bridging the gap between human and machine translation. *arXiv preprint arXiv:1609.08144*, 2016.