Deep equilibrium networks are sensitive to initialization statistics

Atish Agarwala [1] **Samuel S. Schoenholz** [1]

Abstract

Deep equilibrium networks (DEQs) are a promising way to construct models which trade off memory for compute. However, theoretical understanding of these models is still lacking compared to traditional networks, in part because of the repeated application of a single set of weights. We show that DEQs are sensitive to the higher order statistics of the matrix families from which they are initialized. In particular, initializing with orthogonal or symmetric matrices allows for greater stability in training. This gives us a practical prescription for initializations which allow for training with a broader range of initial weight scales.

1. Introduction

Deep equilibrium networks (DEQs) are a network architecture which uses implicitly-defined layers to get benefits of deep networks with a smaller memory footprint (Bai et al., 2019). DEQs have shown competitive performance on image tasks (Bai et al., 2020; Gilton et al., 2021), as well as on language tasks with transformer DEQ models (Bai et al., 2019; 2021).

A DEQ layer can be defined as follows: given an input $\mathbf{x}$, the output $\mathbf{z}^*$ of a DEQ layer is given implicitly by

$$\mathbf{z}^* = f_{\boldsymbol{\theta}}(\mathbf{x}, \mathbf{z}^*) \tag{1}$$

for a function f parameterized by $\boldsymbol{\theta}$. One way to solve for $\mathbf{z}^*$ is via direct iteration; in this case the DEQ can be interpreted as a deep network with identical parameters for each layer (tied weights), as opposed to the more traditional case of independent parameters for each layer (untied weights).

Despite their implicit definition, DEQs can be trained with backpropagation. Equation 1 gives an implicit equation for any Vector-Jacobian Products (VJPs) of the DEQ layer which can also be found using a fixed point solver. Training DEQs requires careful consideration of both initialization as well as the solver used to find the fixed points (Bai et al., 2019; 2021). This gives another way to trade off memory and compute, but brings the additional complication of maintaining convergence of solvers throughout learning. For example, most practical applications use small initializations to maintain stability (Bai et al., 2019). Empirical approaches to stabilize learning include regularization of the Jacobian of the DEQ map (Bai et al., 2021).

While deep networks with untied weights have many theoretically-motivated initialization schemes (Glorot & Bengio, 2010; He et al., 2015; Martens et al., 2021; Schoenholz et al., 2017; Xiao et al., 2018), understanding initialization in implicitly defined networks is a relatively new field (El Ghaoui et al., 2021; Massaroli et al., 2020; Winston & Kolter, 2021). Some progress has been made on understanding DEQs in the wide-network limit by formally relating their properties to wide networks in the untied weights setting (Feng & Kolter, 2020). However, in practical settings DEQs are often trained in a regime where the Jacobians approach divergence (Bai et al., 2019; 2021) - or equivalently, in regimes where the DEQ effectively has large depth. This is precisely the regime where the wide network limit breaks down (Hanin & Nica, 2019).

In this paper we develop a theory of DEQs at initialization which allows us to understand the differences between the DEQ setting and the traditional untied weights setting. In particular, we show that the choice of initial matrix family can lead to quantitatively different behavior even in the wide network limit. Using the linear DEQ, whose output can be computed analytically, we prove the following:

- We compute the convergence properties of the fixed point for random i.i.d. matrices.
- We prove that the tied weights setting is qualitatively and quantitatively different than the untied weights setting, particularly when using non-i.i.d Gaussian weight matrices.
- We show that variance is reduced when initializing with

[1]Google Research, Brain Team. Correspondence to: Atish Agarwala <thetish@google.com>, Samuel S. Schoenholz <schsam@google.com>.

Proceedings of the 39^{th} International Conference on Machine Learning, Baltimore, Maryland, USA, PMLR 162, 2022.

orthogonal or Gaussian orthogonal ensemble (GOE) matrices.

We then examine the convergence properties of a simple non-linear DEQ:

- We derive a self-consistent set of equations which determine the stability of the fixed point.
- We prove that the untied weights theory gives a good approximation to convergence for orthogonal and randomly initialized matrices away from the divergence threshold.
- We show that in the symmetric case the untied weights theory doesn't hold.

We conclude by demonstrating a practical consequence of the theory: networks initialized with random orthogonal or symmetric weight matrices have more stable learning which is likely to converge more quickly than with standard i.i.d. initialization. This increased stability also allows for networks to be trained with a broader set of initial weight scales. This opens a new avenue for DEQ initialization: to optimizing the initialization family rather than just the initialization scale.

2. Theory of linear DEQs

2.1. Fixed point properties

We define a *linear DEQ* $\mathbf{z}^*(\mathbf{x})$ as

$$\mathbf{z}^* = \mathbf{W}\mathbf{z}^* + \mathbf{x} \tag{2}$$

where $\mathbf{x}$ is an N-dimensional input vector and $\mathbf{W}$ is an $N \times N$ matrix. One way to find $\mathbf{z}^*$ is as the limit of the iterated map

$$\mathbf{z}_{t+1} = \mathbf{W}\mathbf{z}_t + \mathbf{x}. \tag{3}$$

This is a linear neural network with input injection at each layer, as well as the same $\mathbf{W}$ shared across layers. Alternatively, we can directly solve for $\mathbf{z}^*$ using the implicit equation:

$$\mathbf{z}^* = (\mathbf{I} - \mathbf{W})^{-1}\mathbf{x}. \tag{4}$$

While this solution exists for all $\mathbf{W}$ with $\mathbf{I} - \mathbf{W}$ invertible, the iterated map is unstable if $\mathbf{W}$ has any eigenvalues with magnitude greater than or equal to 1, as can be seen by the power series

$$\mathbf{z}^* = \lim_{t\to\infty} \mathbf{z}_t = \lim_{t\to\infty} \sum_{k=0}^{t} \mathbf{W}^k\mathbf{x}. \tag{5}$$

A similar convergence criterion was described in (Gao et al., 2022). Convergence can be guaranteed by selecting $\mathbf{W}$ from a parameterized family of matrices with bounded spectral norm, as seen in (Kawaguchi, 2020).

Previous work has focused analysis of the *untied weights* DEQ (Feng & Kolter, 2020). The untied weights DEQ is defined by the iterative equation

$$\mathbf{z}_{t+1} \equiv \mathbf{W}_k\mathbf{z}_t + \mathbf{x} \tag{6}$$

where the $\mathbf{W}_k$ are independent and drawn from the same distribution. In the untied weights case, there is no limit $\mathbf{z}_\infty$; however, the elements $\mathbf{z}_t$ have a common limiting distribution for large t and large N. We will compare the normal linear DEQ (*tied weights* case) with the untied weights case throughout the remainder of this section.

2.1.1. Random Gaussian initialization

We now return to the weight tied setting. For i.i.d. $\mathbf{W}$ with Gaussian entrie, for fixed t the elements of $\mathbf{z}_t$ converge in distribution to Gaussians as $N \to \infty$ with previously derived statistics (Feng & Kolter, 2020; Yang, 2021). The moments are identical to the untied weights case, and, for $\mathbf{z}_0 = 0$ are given by

$$\mathrm{E}[(\mathbf{z}_t)_i] = \mathbf{x}_i,\ \mathrm{Var}[(\mathbf{z}_t)_i] = \frac{1}{N}\mathbf{x}\cdot\mathbf{x}\sum_{k=1}^{t} V^t \tag{7}$$

where the elements of $\mathbf{W}$ (and $\mathbf{W}_t$) have variances V/N.

It has been argued (Feng & Kolter, 2020) that the behavior as $t \to \infty$ can be derived from the $N \to \infty$ limit, by applying the tensor program (TP) framework from (Yang, 2021) (taking the limit $N \to \infty$), and then taking the limit $t \to \infty$. However, the TP framework is well-defined only for programs with finite length. In particular, there is a non-zero probability that the iterated map will not converge even for $V < 1$ due to eigenvalue fluctuations. This has practical consequences as we will show in Section 2.2.

Nevertheless, in the case of linear networks, we are able to prove that $\mathbf{z}^*$ does in fact converge to a Gaussian in distribution in the limit of infinite depth (Appendix A.2), extending the finite-depth result of (Feng & Kolter, 2020). The basic idea of the proof is that, with probability close to 1, $\mathbf{z}_t$ and $\mathbf{z}^*$ become arbitrarily close for large enough N and t if $V < 1$. The convergence in distribution of $\mathbf{z}_t$ can then be used to show convergence in distribution of $\mathbf{z}^*$. This gives us the moments

$$\mathrm{E}[(\mathbf{z}^*)_i] = \mathbf{x}_i,\ \mathrm{Var}[(\mathbf{z}^*)_i] = \frac{1}{N}\mathbf{x}\cdot\mathbf{x}\frac{V}{1-V} \tag{8}$$

as conjectured by (Feng & Kolter, 2020). We provide an alternate derivation of the moments using operator-valued free probability theory in Appendix C.2.

These results require that $\mathbf{W}$ (or the $\mathbf{W}_k$ in the untied weights case) have i.i.d. random entries. In the untied

weights case, for low order moments we can relax the Gaussianity assumption and instead just assume that the $\mathbf{W}_k$ are drawn independently from a rotationally invariant ensemble (distribution over matrices) - extending the result of (Feng & Kolter, 2020) to other matrix families. Let tr define the N-normalized trace (average eigenvalue). Then, if $V = \mathrm{tr}[\mathbf{W}_k^{\mathrm{T}}\mathbf{W}_k] < 1$, $\lim_{t\to\infty} \mathrm{Var}[(\mathbf{z}_t)_i] = \frac{1}{N}\mathbf{x}\cdot\mathbf{x}\frac{V}{1-V}$ (Appendix A.3).

2.1.2. Tied weights, non-Gaussian initialization

However, in the tied weights case, even low order moments can depend on the choice of ensemble for $\mathbf{W}$. We will consider two alternative ensembles:

- *Orthogonal* - $\mathbf{W} = \sqrt{V}\mathbf{O}$ for $\mathbf{O}$ random orthogonal (Haar distributed).
- *GOE* - Random symmetric matrix, Gaussian entries with variance V/N off-diagonal and $2V/N$ on-diagonal.

For DEQs, a basic calculation (Appendix A.3) gives us

$$\mathrm{Var}[\mathbf{z}_i^*] = \frac{1}{N}\left(\mathrm{E}[\mathbf{z}^*\cdot\mathbf{z}^*] - \mathrm{E}[\mathbf{z}^*]\cdot\mathrm{E}[\mathbf{z}^*]\right) \tag{9}$$

The second term evaluates to $\mathbf{x}\cdot\mathbf{x}$. The first term can be written as a power series:

$$\mathrm{E}[\mathbf{z}^*\cdot\mathbf{z}^*] = \mathrm{E}_{\mathbf{W}}\left[\mathrm{tr}\left[\left(\sum_{k=0}^{\infty}\mathbf{W}^k\right)^{\mathrm{T}}\left(\sum_{k=0}^{\infty}\mathbf{W}^k\right)\right]\right]\mathbf{x}\cdot\mathbf{x} \tag{10}$$

which converges for V less than some V_c. In the i.i.d. random case, $V_c = 1$.

We evaluate the power series in Appendix A.3. In the orthogonal case, the variance matches the i.i.d. random case, with $V_c = 1$. However the GOE case gives dramatically different behavior. We can compute the variance using Equation 10. We note that $\mathbf{W}_k^{\mathrm{T}} = \mathbf{W}_k$, and that $\mathrm{tr}[\mathbf{W}^k] = C_k V^k$, where C_k is the kth Catalan number. Using generating series, we have

$$\mathrm{Var}[(\mathbf{z}_i^*)^2] = \frac{1}{\sqrt{1-4V}} - \frac{1-\sqrt{1-4V}}{2V} - 1 \tag{11}$$

One interesting feature is that $\mathrm{Var}[(\mathbf{z}_i^*)^2]$ diverges differently for symmetric $\mathbf{W}$ than for random and orthogonal $\mathbf{W}$. The divergence happens at $V_c = 1/4$ - which reflects the spectral radius of $2\sqrt{V}$ of the semi-circular law.

The behavior near V_c is of interest as well. For $\delta \equiv 1 - V/V_c$, for the GOE ensemble the asymptotic behavior of the variance of $(\mathbf{z}_i^*)^2$ is given by $O(\delta^{-0.5})$ for $\delta \ll 1$, while for random and orthogonal the divergence is $O(\delta^{-1})$. As we will see, the differences become more pronounced for higher order moments.

The difference in $\mathrm{Var}[(\mathbf{z}_i^*)^2]$ is relevant for learning dynamics. This can be seen explicitly in the large width limit where the neural tangent kernel (NTK) controls learning (Jacot et al., 2018). For the linear DEQ, the NTK is given by $\frac{1}{N^2}\mathrm{E}[\mathbf{z}^*\cdot\mathbf{z}^*]^2\mathbf{x}\cdot\mathbf{x}'$ for an input pair $(\mathbf{x},\mathbf{x}')$ (Appendix B.1) - so the GOE and random ensembles give different functions in the wide network limit.

2.2. Variability of outputs

In the limit of infinite width, the first and second moments of $\mathbf{z}^*$ are often sufficient to characterize the output of the DEQ. We saw that the first and second moments of $\mathbf{z}^*$ are identical for random and orthogonal DEQs with tied weights, which match the statistics of the untied weights case. However, the GOE ensemble had a different second moment. This already suggests that different distributions for $\mathbf{W}$ have different behavior.

In the wide but finite dimensional setting, the differences between the matrix families become more stark. While the full distribution for finite N is intractable, we can understand some of the differences by understanding a particular 4th moment of $\mathbf{z}^*$. These differences will be reflected in the convergence dynamics of $\mathbf{z}^*$, as well as the implicit bias of the DEQ.

We begin by defining a particular 4th moment which we call the *length variance*. Given a random input $\mathbf{x}$ and a random matrix $\mathbf{M}$ we have:

Definition 2.1. For $\mathbf{z} = \mathbf{M}\mathbf{x}$, the *length variance* of $\mathbf{z}$ is defined as:

$$\sigma^2_{\mathbf{z}\cdot\mathbf{z}} = \frac{1}{N}\mathrm{E}_{\mathbf{M}}[\mathrm{Var}_{\mathbf{x}}[\mathbf{z}\cdot\mathbf{z}]] \tag{12}$$

In order to compute the length variance, we use the following lemma (proof in Appendix A.4):

Lemma 2.2. *Let $\mathbf{M}$ be an $N\times N$-dimensional random matrix. Let $\mathbf{x}$ be an N-dimensional vector with i.i.d. Gaussian elements with 0 mean and variance $\sigma^2_{\mathbf{x}}$, and let $\mathbf{z} = \mathbf{M}\mathbf{x}$. Then we have:*

$$\sigma^2_{\mathbf{z}\cdot\mathbf{z}} = \frac{2}{N}\sigma^4_{\mathbf{x}}\mathrm{E}_{\mathbf{M}}\left[\mathrm{tr}[(\mathbf{M}^{\mathrm{T}}\mathbf{M})^2]\right] \tag{13}$$

Without loss of generality, we assume that $\sigma^2_{\mathbf{x}} = 1$ for the remainder of this section.

2.2.1. Untied weights case

We first consider the case of untied weights, where we compute $\sigma^2_\infty \equiv \lim_{t\to\infty}\sigma^2_{\mathbf{z}_t\cdot\mathbf{z}_t}$. In order to define $\sigma^2_{\mathbf{z}_t\cdot\mathbf{z}_t}$ in terms of Lemma 2.2, $\mathbf{x}$ is the random input, and $\mathbf{M} = \sum_{k=0}^{t}\prod_{j=0}^{k}\mathbf{W}_k$.

In the random and GOE cases, a direct calculation (Appendix A.4) shows that in the limit of large N we have

$$\sigma_\infty^2 = \frac{2}{N}\left(\frac{2}{(1-V)^2} + \frac{1}{(1-V^2)^2} - \frac{2}{1-V^2}\right) \quad (14)$$

Meanwhile, for orthogonal $\mathbf{W}_k$ we have

$$\sigma_\infty^2 = \frac{2}{N}\left(\frac{2}{(1-V)^2} - \frac{1}{1-V^2}\right) \quad (15)$$

The two main features here are that σ_∞^2 is well-defined in the untied weights case, and for $\delta \equiv 1 - V$, $\sigma_\infty^2 = O(\delta^{-2})$.

2.2.2. Tied weights case

The tied weights case is different. We want to compute

$$\sigma^2_{\mathbf{z}^*\cdot\mathbf{z}^*} = \frac{2}{N}\mathrm{E}_{\mathbf{W}}\left[\mathrm{tr}\left[((\mathbf{I}-\mathbf{W})^{-\mathrm{T}}(\mathbf{I}-\mathbf{W})^{-1})^2\right]\right] \quad (16)$$

We first attempt solution via the formal power series

$$\sigma^2_{\mathbf{z}^*\cdot\mathbf{z}^*} = \frac{2}{N}\mathrm{E}_{\mathbf{W}}\left[\mathrm{tr}\left[\sum_{j,k,l,m=0}^{\infty}(\mathbf{W}^j)^{\mathrm{T}}\mathbf{W}^k(\mathbf{W}^l)^{\mathrm{T}}\mathbf{W}^m\right]\right] \quad (17)$$

In the orthogonal weights case the power series converges for $V < 1$ (Appendix A.4):

$$\sigma^2_{\mathbf{z}^*\cdot\mathbf{z}^*} = \frac{2}{(1-V)^3} - \frac{1}{(1-V)^2} \quad (18)$$

which scales as $O(\delta^{-3})$ for $\delta \equiv 1 - V$. We immediately see that there is more variance in the tied weights setting.

However, for random and GOE matrices the power series in Equation 17 diverges with non-zero probability. This is due to the previously mentioned fluctuations in the largest eigenvalues, which mean that iteration can diverge in the random and GOE cases even when $V < V_c$. Therefore we attempt to compute the right hand side of Equation 16 directly in the large N limit. This can be done using operator-valued free probability theory (Appendix C.2).

In the random case we have

$$\lim_{N\to\infty} N\sigma^2_{\mathbf{z}^*\cdot\mathbf{z}^*} = \frac{2V^2}{(1-V)^4} + \frac{4V}{(1-V)^3} + \frac{2}{(1-V)^2} \quad (19)$$

in the limit of large N. For $\delta \equiv 1 - V$ we get $O(\delta^{-4})$ for small δ. In the GOE case, we have

$$\lim_{N\to\infty} N\sigma^2_{\mathbf{z}^*\cdot\mathbf{z}^*} = \frac{1}{2V}\left(\frac{1}{(1-4V)^{-5/2}} - \frac{1}{(1-4V)^{-3/2}}\right) \quad (20)$$

which diverges as $O(\delta^{-2.5})$ for $\delta \equiv 1 - 4V$.

2.2.3. Divergence of the variance

The most interesting feature of $\sigma^2_{\mathbf{z}^*\cdot\mathbf{z}^*}$ is its behavior as V approaches the critical threshold V_c, where the iterative solving fails. As $\delta \equiv 1 - V/V_c$ approaches 0, $\sigma^2_{\mathbf{z}^*\cdot\mathbf{z}^*}$ diverges. This reflects the increased variability in the outputs - both within a single DEQ and between DEQ models. This can lead to instability in learning.

In the untied weights case the divergence goes as $O(\delta^{-2})$ irrespective of the matrix ensemble. However, in the tied weights case we have different, and worse, behavior ($O(\delta^{-2.5})$ for GOE, $O(\delta^{-3})$ for orthogonal, and $O(\delta^{-4})$ for i.i.d. random). For the random and GOE, these behaviors are valid when N is large enough that the fluctuations in the largest eigenvalues are small ($N \gg \delta^{-2/3}$, when Tracy-Widom fluctuations (Tracy & Widom, 1994) are controlled); there will be even more variability when $N \ll \delta^{-2/3}$. This is why the infinite power series solution fails; the direct calculation involving $(\mathbf{I} - \mathbf{W})^{-1}$, as N goes to infinity, is the equivalent of taking the limit $\lim_{t\to\infty}\lim_{N\to\infty}\sigma^2_{\mathbf{z}_t\cdot\mathbf{z}_t}$.

We can confirm these results numerically by computing $\mathrm{tr}[[(\mathbf{I} - \sqrt{V}\mathbf{W})^{\mathrm{T}}(\mathbf{I} - \sqrt{V}\mathbf{W})]^{-2}]$ for various V for a fixed $\mathbf{W}$, drawn separately for each of the distributions. We plot the trace against δ (Figure 1).

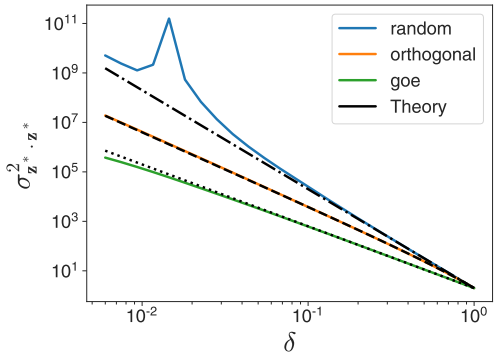

Figure 1. Empirical length variance $\sigma^2_{\mathbf{z}^*\cdot\mathbf{z}^*}$ computed using $\mathrm{tr}[[(\mathbf{I}-\sqrt{V}\mathbf{W})^{\mathrm{T}}(\mathbf{I}-\sqrt{V}\mathbf{W})]^{-2}]$ for $N = 5000$. Plotted against $\delta = 1 - V/V_c$ ($V_c = 1/4$ for GOE and 1 otherwise). Orthogonal is well predicted by theory, which also predicts intermediate scale behavior for GOE and i.i.d. random. Random shows the most variance, including non-monotonic behavior due to eigenvalue fluctuations.

As expected, for the orthogonal $\mathbf{W}$ the trace follows its theoretical distribution due to self-averaging in large dimensions. We also see that the GOE has a smaller trace than the orthogonal, but begins to deviate from the infinite-width limit for small δ. Finally we see that the random case tends to have larger variability than either the random or orthogonal. The non-monotonicity in this example is associated with the eigenvalue with largest real part approaching 1, even when

$V < V_c$, due to finite-size fluctuations.

This analysis suggests that in practice, the orthogonal and GOE ensembles give more stable performance when initialized near the transition. The orthogonal ensemble is guaranteed to converge when $V < 1$. While the GOE ensemble is also affected by fluctuations which can cause eigenvalues larger than 1, the smaller density of states at the edge of the distribution means it is more stable to fluctuations than the random ensemble.

3. Theory of non-linear DEQs

3.1. Basic model and statistics

Many of the ideas from the linear DEQ are reflected in the non-linear setting as well. We consider the non-linear DEQ:

$$\mathbf{z}^* = \phi(\mathbf{W}\mathbf{z}^*) + \mathbf{x} \tag{21}$$

where ϕ is an elementwise non-linearity. In order to understand the statistics in the wide network limit, it is useful to instead study $\mathbf{h}^* = \mathbf{W}\mathbf{z}^*$ defined by

$$\mathbf{h}^* = \mathbf{W}\phi(\mathbf{h}^*) + \mathbf{W}\mathbf{x} \tag{22}$$

As with the linear DEQ, $\mathbf{h}^*$ can be found as the limit of the iterated map

$$\mathbf{h}_{t+1} = \mathbf{W}\phi(\mathbf{h}_t) + \mathbf{W}\mathbf{x} \tag{23}$$

For random $\mathbf{W}$, the iterated map implied by Equation 23 limits to a Gaussian process for a fixed number of iterations in the large N limit (Feng & Kolter, 2020; Yang, 2021). It remains an open question to show that as the number of iterations increases, there is a limiting GP whose properties can be computed by solving a fixed point equation for the kernel. If such a limiting kernel exists, it can be solved for using the methods from (Feng & Kolter, 2020).

Following previous work on mean field theory of neural networks (Feng & Kolter, 2020; Lee et al., 2019; Poole et al., 2016; Schoenholz et al., 2017), we study $\sigma^2_{\mathbf{h}^*} \equiv \mathrm{E}_{\mathbf{W}}\left[\frac{1}{N}\mathbf{h}^* \cdot \mathbf{h}^*\right]$ in the wide network limit. In the GP limit, $\sigma^2_{\mathbf{h}^*} = \mathrm{Var}[\mathbf{h}^*_i]$; however, we can study $\sigma^2_{\mathbf{h}^*}$ explicitly for large N even when the final limiting distribution is unknown.

We note that $\sigma^2_{\mathbf{h}^*}$ obeys the following self-consistent equation:

$$\sigma^2_{\mathbf{h}^*} = \mathrm{E}_{\mathbf{W}}\left[\frac{1}{N}(\phi(\mathbf{h}^*) + \mathbf{x})^{\mathrm{T}}\mathbf{W}^{\mathrm{T}}\mathbf{W}(\phi(\mathbf{h}^*) + \mathbf{x})\right] \tag{24}$$

In order to solve for $\sigma^2_{\mathbf{h}^*}$, we need to make some assumptions about $\mathbf{h}^*$, $\mathbf{x}$, and $\mathbf{W}$. We can prove the following:

Theorem 3.1. *Suppose $\phi(\mathbf{h}^*)$, $\mathbf{x}$, and $\mathbf{W}$ are freely independent, and suppose that the distribution of $\mathbf{W}$ is rotationally symmetric. Then we have:*

$$\sigma^2_{\mathbf{h}^*} = V\left(\sigma^2_{\phi(\mathbf{h}^*)} + 2C_{\mathbf{x},\phi(\mathbf{h}^*)} + \sigma^2_{\mathbf{x}}\right) \tag{25}$$

where $V = \mathrm{E}[\mathrm{tr}[\mathbf{W}\mathbf{W}^{\mathrm{T}}]]$, $\sigma^2_{\mathbf{x}} \equiv \frac{1}{N}\mathbf{x} \cdot \mathbf{x}$ and

$$\sigma^2_{\phi(\mathbf{h}^*)} \equiv \mathrm{E}[\phi(\mathbf{h}^*_i)^2],\ C_{\mathbf{x},\phi(\mathbf{h}^*)} \equiv \left(\frac{1}{N}\mathbf{1}^{\mathrm{T}}\mathbf{x}\right)\mathrm{E}[\phi(\mathbf{h}^*_i)] \tag{26}$$

If $\mathbf{W}$ is a random orthogonal matrix (times a fixed scale), then Equation 25 holds as long as $\phi(\mathbf{h}^*)$ and $\mathbf{x}$ are freely independent.

Proof. Using free independence of $\mathbf{W}$ with respect to the other variables we have

$$\sigma^2_{\mathbf{h}^*} = \mathrm{E}_{\mathbf{W}}\left[\mathrm{tr}\left[\mathbf{W}\mathbf{W}^{\mathrm{T}}\right]\right]\mathrm{E}_{\mathbf{W}}\left[\frac{1}{N}||\phi(\mathbf{h}^*) + \mathbf{x}||^2\right] \tag{27}$$

which evaluates to

$$\sigma^2_{\mathbf{h}^*} = V\left(\mathrm{E}_{\mathbf{W}}\left[\mathrm{tr}\left[(\phi(\mathbf{h}^*)\phi(\mathbf{h}^*)^{\mathrm{T}} + 2\phi(\mathbf{h}^*)\mathbf{x}^{\mathrm{T}} + \mathbf{x}\mathbf{x}^{\mathrm{T}}\right]\right]\right) \tag{28}$$

Using the independent of $\mathbf{x}$ and $\phi(\mathbf{h}^*)$, we arrive at Equation 25. If $\mathbf{W}$ is a random orthogonal matrix, then Equation 28 follows directly from Equation 24, and the rest of the analysis holds.

□

In Figure 2, we iteratively solve for $\mathbf{h}^*$ for $\mathbf{W}$ using the different matrix families with different scales. We then numerically solve Equation 25. Comparing the theoretical $\sigma^2_{\mathbf{h}^*}$ with the empirical $\frac{1}{N}\mathbf{h}^* \cdot \mathbf{h}^*$ (for $N = 1000$), we see that for the random and orthogonal families, the empirical and theoretical quantities are well matched. The agreement in the orthogonal case is not surprising, as the $\mathbf{W}^{\mathrm{T}}\mathbf{W}$ in Equation 27 evaluates to the identity matrix.

However, for the GOE, the actual variance is higher than the predicted one. This suggests that $\phi(\mathbf{h}^*)$, $\mathbf{x}$, and $\mathbf{W}$ are not freely independent in the symmetric case. In the linear case this was due to the fact that the left and right singular vectors of $\mathbf{W}$ are identical in the GOE case. Here the presence of the non-linearity ϕ prevents the application of the linear theory, but the difference between the GOE and the other ensembles persists.

3.2. Fixed point stability

For non-linear DEQs, there is no general analytic solution for $\mathbf{z}^*$ or $\mathbf{h}^*$. However, one can still understand the stability of fixed points of the iterative map of Equation 23. Linearizing the dynamics in $\delta\mathbf{h}_t \equiv \mathbf{h}_t - \mathbf{h}^*$, for small differences

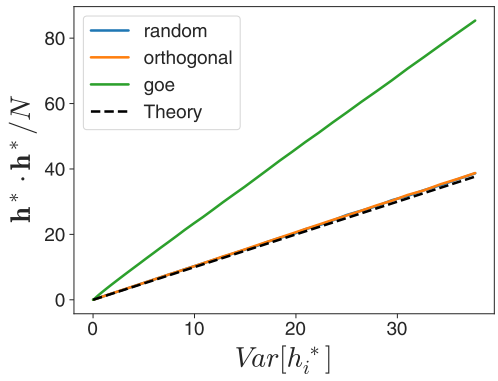

Figure 2. Empirical $\frac{1}{N}\mathbf{h}^* \cdot \mathbf{h}^*$ ($N = 1000$) versus theoretical $\mathrm{Var}[(\mathbf{h}_i^*)^2]$ for ϕ hard-tanh. Random and orthogonal cases coincide and are well-predicted by the assumption of $\mathbf{h}$ and $\mathbf{W}$ while GOE variance is larger.

we have:

$$\delta\mathbf{h}_{t+1} = \mathbf{W} \circ \phi'(\mathbf{h}^*)\delta\mathbf{h}_t + \mathcal{O}(\delta\mathbf{h}^2) \tag{29}$$

where $\circ\phi'(\mathbf{h}^*)$ represents the Hadamard product - multiplication by a diagonal matrix with entries $\phi'(\mathbf{h}^*)$.

The fixed point is stable under iteration if and only if $\mathbf{W} \circ \phi'(\mathbf{h}^*)$ has eigenvalues with absolute value less than 1. We expect that the self-consistent Equation 24 to hold near any fixed point. Indeed, Equation 24 also describes the situation where there is no stable fixed point but the statistics of iteration converge. Therefore, by computing the empirical limiting value of $\frac{1}{N}\mathbf{h}^* \cdot \mathbf{h}^*$, we can predict the existence of a stable fixed point. If the statistics of $\mathbf{h}^*$ are consistent, then all such fixed points will be stable.

We must posit a relationship between $\mathbf{W}$ and $\phi'(\mathbf{h}^*)$ to compute the maximum eigenvalue. For random and orthogonal $\mathbf{W}$, we have the following theorem:

Theorem 3.2. *Let $\mathbf{W}$ be an $N \times N$ matrix and $\mathbf{h}^*$ be an $N \times 1$ dimensional vector. Let $\mathbf{W}$ and $\phi'(\mathbf{h}^*)$ be freely independent. If $\mathbf{W}$ is a random Gaussian matrix or a random orthogonal matrix, as $N \to \infty$ the spectral radius r of $\mathbf{W} \circ \phi'(\mathbf{h}^*)$ is given by*

$$r^2 = \mathrm{E}_{\mathbf{W}}[\mathrm{tr}[\mathbf{W}^{\mathrm{T}}\mathbf{W}]]\mathrm{E}\left[\frac{1}{N}\phi'(\mathbf{h}^*) \cdot \phi'(\mathbf{h}^*)\right] \tag{30}$$

Proof. In order to prove the theorem, we will prove that $\mathbf{W} \circ \phi'(\mathbf{h}^*)$ are *R-diagonal* operators - which have rotationally symmetric empirical spectra in the complex plane with known maximum radius (Mingo & Speicher, 2017). We will use the characterization from (Nica & Speicher, 1996), where an element $\mathbf{R}$ is R-diagonal if it obeys a polar decomposition

$$\mathbf{R} = \mathbf{UP} \tag{31}$$

where the sets $\{\mathbf{U}, \mathbf{U}^\dagger\}$ and $\{\mathbf{P}, \mathbf{P}^\dagger\}$ are freely independent, and $\mathbf{U}$ is Haar distributed. We note that $\mathbf{R}$ has a Brown measure with support given by an annulus in the complex plane centered at zero, with maximum radius $r^2 = \mathrm{E}[\mathrm{tr}[\mathbf{R}^\dagger\mathbf{R}]]$.

It remains to show that $\mathbf{W} \circ \phi'(\mathbf{h}^*)$ is $\mathbf{R}$-diagonal. If $\mathbf{W}$ is orthogonal, we already have a polar decomposition. If $\mathbf{W}$ is a Gaussian random matrix, we can use the singular value decomposition $\mathbf{W} = \mathbf{U}\boldsymbol{\sigma}\mathbf{V}$. We have $\mathbf{U} = \mathbf{U}$ and $\mathbf{P} = \boldsymbol{\sigma}\mathbf{V} \circ \phi'(\mathbf{h}^*)$. We immediately have that $\{\mathbf{U}, \mathbf{U}^\dagger\}$ and $\{\mathbf{P}, \mathbf{P}^\dagger\}$ are freely independent, and $\mathbf{R}$ is R-diagonal.

Computing the trace of $\mathbf{R}^{\mathrm{T}}\mathbf{R}$ completes the proof. □

The GOE case is more complicated even when $\phi'(\mathbf{h}^*)$ and $\mathbf{W}$ are freely independent. Progress can be made if ϕ is monotonic. In this case, $\phi'(\mathbf{h}^*)$ is non-negative, and $\mathbf{W} \circ \phi'(\mathbf{h}^*)$ has the same spectrum as the symmetric matrix $\phi'(\mathbf{h}^*)^{1/2} \circ \mathbf{W} \circ \phi'(\mathbf{h}^*)^{1/2}$. Free probability theory can then be used to numerically compute the spectral radius of $\mathbf{W} \circ \phi'(\mathbf{h}^*)$ (Mingo & Speicher, 2017; Nica & Speicher, 1996).

In certain cases, we can compute the spectral radius analytically. If ϕ is the hard-tanh function, then the distribution is a linear combination of a δ-function at 0 and a semicircular law (Appendix C.1). The spectral radius is given by $2\sqrt{Vp(\mathbf{h}^*)}$, where $p(\mathbf{h}^*)$ is the probability that $|\mathbf{h}_i^*| < 1$ for $\mathbf{h}_i^*$ Gaussian distributed with some variance $\sigma^2_{\mathbf{h}^*}$ and mean 0. Up to a factor of 2 this is equivalent to the radius for freely-independent orthogonal and random $\mathbf{W}$.

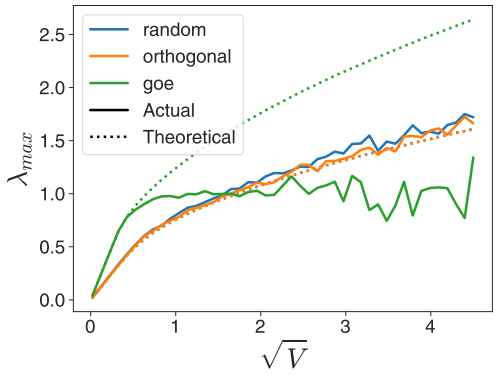

Figure 3. Theoretical versus actual λ_{max} for DEQ with hard-tanh non-linearity. Free-probability theory predicts random and orthogonal λ_{max} for $\lambda_{max} \leq 1$. GOE is well predicted for small $\sqrt{V}$ but not near the transition.

We can compare these theoretically-predicted stability conditions with stability in practice. Given a randomly initialized DEQ, we can compute the largest eigenvalue (by absolute value) and compare it to the theoretical prediction for different values of $\sqrt{V}$ (Figure 3). We see that for small V, all matrix families are well-described by the

theory. However, in the symmetric case we see deviations at intermediate V as well. As λ_{max} approaches 1 we see more deviations; finally, when $\lambda_{max} > 1$ there are large fluctuations in the maximum eigenvalue.

For the random and orthogonal cases, the theoretical curves can be used to predict the transition from a stable fixed point to no stable fixed point. One way to detect convergence is to plot the residual $||\mathbf{h}_{t+1} - \mathbf{h}_t||$ for some large t after forward iteration of the DEQ map. For all matrix families, this residual goes from near-zero to a non-zero value as V increases (Figure 4, averaged over 1000 samples). We can predict the critical $\sqrt{V}$ using the theoretical predictions for λ_{max}. For the orthogonal and random cases, we see that the transition is well-predicted by the theory, while in the symmetric case the theory predicts a transition earlier than the actual transition.

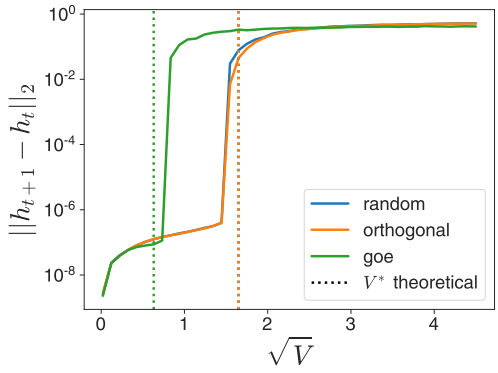

Figure 4. $||\mathbf{h}_{t+1} - \mathbf{h}_t||$ for DEQs with hard-tanh non-linearity ($N = 1000$, averaged over 1000 samples). For large $\sqrt{V}$, DEQ doesn't converge. Divergence threshold is well predicted for random and orthogonal, but more variance for GOE ensemble.

Even in the non-linear case the random and orthogonal families have similar large-width behavior (which is predictive of some finite width properties), but the symmetric matrix has different properties due to its normality.

4. Experiments

The theoretical analysis has shown that the different matrix families have different properties in both the wide network and finite width limits. In particular, the orthogonal and GOE ensembles often have less variability than the random ensemble. In this section, we empirically explore the effects of initializing practical models with different matrix families. We start by showing that orthogonal initializations increase stability and performance for DEQs trained on MNIST. We then show that for DEQ transformers, using orthogonal or GOE ensembles increases the stability, and sometimes even the speed of learning compared to i.i.d. random matrices.

4.1. Fully connected DEQ on MNIST

We begin by training a fully-connected DEQ on MNIST (LeCun et al., 1989). The network consists of a flattening layer, followed by a ResNet DEQ layer with tanh non-linearity, and finally a dense layer to project to 10 logits.

We initialize the DEQ layer with random and orthogonal weight matrices at different scales. Learning rate tuning of an ADAM optimizer with momentum of 0.9 suggested an optimal learning rate of 10^{-2} for all the conditions studied. We then trained networks to convergence with 10 random seeds for each initialization family-scale pair.

For small scales, the test error is similar for both types of initializations (Figure 5). However, for larger scales, the orthogonal initialization obtains lower test error. The learning for the random initialization is less stable, as evidenced by a gap between the median and mean test error across the random seeds. This suggests that the orthogonal initialization increases the volume of hyperparameter space where DEQs are viable.

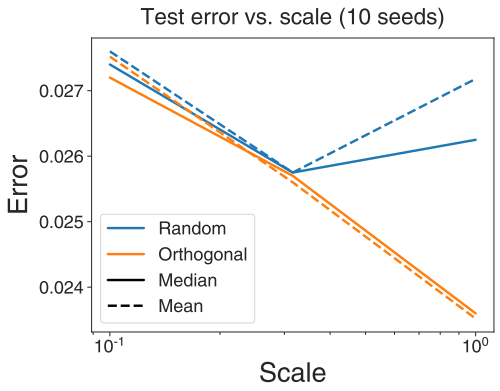

Figure 5. Test error for fully connected DEQs trained on MNIST. Learning rate tuning was performed for each initialization-scale pair, after which statistics were taken over 10 random seeds. Orthogonal initialization allows for larger initialization scales to train stably, and achieves lower test error.

4.2. DEQ transformer

We next examine the effects of the matrix ensembles on a DEQ using a vanilla transformer layer from (Al-Rfou' et al., 2019) as the base of the DEQ layer, trained on Wikitext103 (Merity et al., 2016). With this architecture, there are two main sets of matrices to initialize: the attention matrices and the dense layers. We focused on the dense layers, as they have the same structure as the theoretical calculations.

We modified a Haiku implementation of the DEQ transformer (Khan, 2020). The details of our training procedure can be found in Appendix D. We trained on TPUv3.

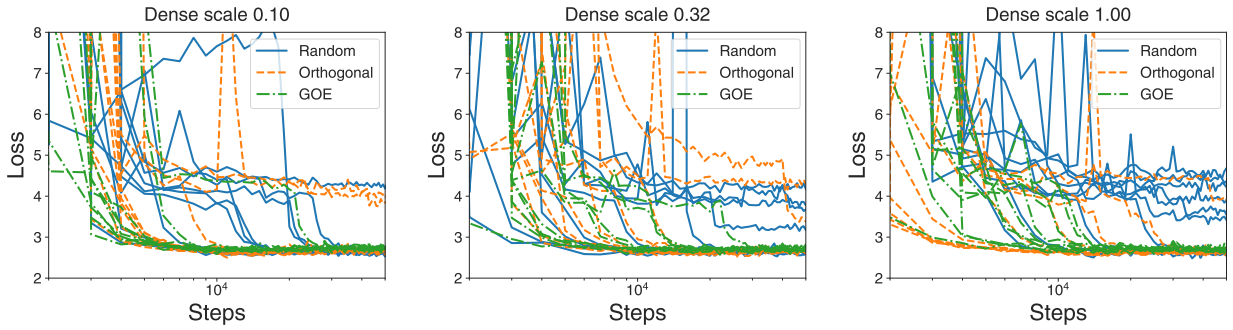

Figure 6. Test loss for various $\sqrt{V}$ and initial matrix ensembles, 10 independent seeds. While random initializations reach lowest test loss, they have a higher chance of diverging, and generally converge more slowly. GOE and orthogonal initializations perform better as $\sqrt{V}$ increases.

4.3. Stability and variability of learning

We trained models with each of three matrix families (random, GOE, and random orthogonal), with $\sqrt{V} \in [10^{-1}, 10^{0}]$ over 10 random seeds. We find that the average test loss is best for the GOE across all $\sqrt{V}$, and orthogonal is better than random at large $\sqrt{V}$ (Table 1). However, we see that comparing the best seeds from each family, the GOE performs worst.

Table 1. Test perplexity on WIkitext103 for different $\sqrt{V}$ and initial matrix ensembles. Min and average taken over 10 random seeds. Random initialization has best performing models, but less stable learning than GOE and orthogonal.

	GOE		ORTHOGONAL		RANDOM	
$\sqrt{V}$	MIN	AVE	MIN	AVE	MIN	AVE
0.1	68.1	71.8	60.7	162.7	56.8	153.9
0.3	66.1	69.8	60.5	173.4	56.3	224.9
1.0	66.3	68.3	57.4	112.1	55.6	481.5

The discrepancy between the average and minimum can be understood by looking at the learning curves themselves. For $\sqrt{V} = 0.1$, where the random family performs best, we see that many trajectories converge slowly to the equilibrium value (Figure 6), and some don't converge at all. In comparison, the orthogonal family has trajectories which don't converge, but those that do converge more quickly. All families from the GOE ensemble converge.

The difference between the families is more dramatic for $\sqrt{V} = 1$. The random initialization fails to converge to low test loss very often compared to the orthogonal and GOE ensembles. This suggests that a broader range of hyperparameters can be stably explored using non-i.i.d. initializations, and that previously reported limits of initialization (Bai et al., 2019; 2020; 2021) may be overcome even without regularization.

This suggests that we can increase the stability of learning by switching to a different matrix family. The GOE is the most stable, but also has the highest minimum perplexity. The orthogonal family is a better choice, as it trades off less aggressively between performance and stability.

5. Conclusions

5.1. Differences in the large-width limit

Our theory and experiments suggested that even in the large width limit, symmetric matrices behave differently from random and orthogonal matrices. Recent work has shown that orthogonal and random matrices have similar behavior for finite depth and large width (Huang et al., 2021; Martens, 2021); we conjecture that the same may be true for DEQs.

In the linear DEQ case, we can understand the similarities and differences by noting that, for random and orthogonal matrices, $\mathrm{tr}[(\mathbf{W}^k)^{\mathrm{T}}\mathbf{W}^j] = \delta_{jk}\mathrm{tr}[\mathbf{W}^{\mathrm{T}}\mathbf{W}]^k$. For i.i.d. random matrices, this is due to the fact that the right and left singular vectors of $\mathbf{W}$ are independent; for the orthogonal family, this is due to the fact that the singular values are all 1.

However the symmetric basis has identical left and right singular vectors, so the singular values of matrix powers are not determined solely by the average (squared) singular value of $\mathbf{W}$ itself. We note that in the untied weights case, this is not an issue; here instead of spectra of $\mathbf{W}^k$, we care about $\prod_{i=1}^{k}\mathbf{W}_i$ for independent $\mathbf{W}_i$, which can be computed using the individual traces.

In general, matrices from *normal* families will behave differently from random i.i.d. matrices. With any normal family, the iterated DEQ map may not display free-independence layer to layer. As we saw in Section 3, this is reflected in the non-linear case as well.

We also saw that for all families, the tied weights case had

asymptotically more finite-width deviations than the equivalent untied weights case as the networks approached the divergence threshold. As many DEQs are eventually trained near the unstable regime (Bai et al., 2019; 2021), these differences are relevant for understanding the performance of real DEQs.

5.2. Practical implications

DEQs suffer from training instability (Bai et al., 2019; 2021), perhaps more so than "traditional" networks due to their (implicitly) iterative nature. One approach is to use alternative parameterizations to induce Lipschitz bounds with respect to parameters (Revay et al., 2020), or to guarantee convergence independent of parameter values (Kawaguchi, 2020; Winston & Kolter, 2021). Another promising approach involves regularization of the Jacobian ($\mathbf{W} \circ \phi'(\mathbf{h}^*)$ in our simple case) to stabilize learning (Bai et al., 2021).

Our theoretical work suggested a different way to to mitigate this instability by optimizing over different matrix ensembles. In the linear, untied weights case (normal deep network), learning dynamics with orthogonal initializations has been extensively studied (Saxe et al., 2014), and orthogonal initializations have been proposed as a method of reducing training instability (Hu et al., 2019; Schoenholz et al., 2017; Xiao et al., 2018).

The orthogonal family provided clear benefits on MNIST, increasing the range of good initialization scales, and eventually leading to better test performance. For the vanilla DEQ transformer, the alternative matrix ensembles provide benefits to training dynamics. While our experiments didn't show significant gains to the best performing networks, the increase in stability and consequently typical training speed was significant. This gives a simple way to improve training setups, and also seems to allow for a broader range of hyperparameters to be explored. With improved tuning, alternate matrix families may be able to improve on best case performance instead of just average case.

5.3. Future directions

Given the success of the GOE ensemble, it may be useful to study other normal ensembles. Random antisymmetric matrices have been used to initialize RNN models (Chang et al., 2018). One can construct a Haar-distributed random normal family with any eigenvalue distribution of interest, complex or real.

Our analysis focused on fixed point solution via naive forward iteration, while in practice (including in our own transformer experiments) quasi-Newton methods like the Broyden solver with Anderson acceleration are used to speed up convergence to the fixed point. It may be possible to analyze these algorithms theoretically as well.

Another possible extension is applying similar methods to other implicitly defined networks like neural ODEs (Chen et al., 2018), which require numerical solution of fixed points. Initializing with different matrix ensembles may improve performance of these methods.

References

Al-Rfou', R., Choe, D., Constant, N., Guo, M., and Jones, L. Character-Level Language Modeling with Deeper Self-Attention. In *AAAI*, January 2019.

Bai, S., Kolter, J. Z., and Koltun, V. Deep Equilibrium Models. In *Advances in Neural Information Processing Systems*, volume 32. Curran Associates, Inc., 2019.

Bai, S., Koltun, V., and Kolter, J. Z. Multiscale Deep Equilibrium Models. *arXiv:2006.08656 [cs, stat]*, November 2020.

Bai, S., Koltun, V., and Kolter, J. Z. Stabilizing Equilibrium Models by Jacobian Regularization. *arXiv:2106.14342 [cs, stat]*, June 2021.

Chang, B., Chen, M., Haber, E., and Chi, E. H. AntisymmetricRNN: A Dynamical System View on Recurrent Neural Networks. In *International Conference on Learning Representations*, September 2018.

Chen, T. Q., Rubanova, Y., Bettencourt, J., and Duvenaud, D. K. Neural Ordinary Differential Equations. In *NeurIPS*, January 2018.

El Ghaoui, L., Gu, F., Travacca, B., Askari, A., and Tsai, A. Implicit Deep Learning. *SIAM Journal on Mathematics of Data Science*, 3(3):930–958, January 2021. doi: 10.1137/20M1358517.

Feng, Z. and Kolter, J. Z. On the Neural Tangent Kernel of Equilibrium Models. September 2020.

Gao, T., Liu, H., Liu, J., Rajan, H., and Gao, H. A global convergence theory for deep ReLU implicit networks via over-parameterization, February 2022.

Gilton, D., Ongie, G., and Willett, R. Deep Equilibrium Architectures for Inverse Problems in Imaging. *arXiv:2102.07944 [cs, eess]*, June 2021.

Glorot, X. and Bengio, Y. Understanding the difficulty of training deep feedforward neural networks. In *Proceedings of the Thirteenth International Conference on Artificial Intelligence and Statistics*, pp. 249–256. JMLR Workshop and Conference Proceedings, March 2010.

Hanin, B. and Nica, M. Finite Depth and Width Corrections to the Neural Tangent Kernel. In *International Conference on Learning Representations*, September 2019.

He, K., Zhang, X., Ren, S., and Sun, J. Delving Deep into Rectifiers: Surpassing Human-Level Performance on ImageNet Classification. In *Proceedings of the IEEE International Conference on Computer Vision*, pp. 1026–1034, 2015.

Hu, W., Xiao, L., and Pennington, J. Provable Benefit of Orthogonal Initialization in Optimizing Deep Linear Networks. In *International Conference on Learning Representations*, September 2019.

Huang, W., Du, W., and Da Xu, R. Y. On the Neural Tangent Kernel of Deep Networks with Orthogonal Initialization. *arXiv:2004.05867 [cs, stat]*, July 2021.

Jacot, A., Gabriel, F., and Hongler, C. Neural Tangent Kernel: Convergence and Generalization in Neural Networks. In *Advances in Neural Information Processing Systems 31*, pp. 8571–8580. Curran Associates, Inc., 2018.

Kawaguchi, K. On the Theory of Implicit Deep Learning: Global Convergence with Implicit Layers. In *International Conference on Learning Representations*, September 2020.

Khan, A. DEQ Jax, 2020.

Kudo, T. and Richardson, J. SentencePiece: A simple and language independent subword tokenizer and detokenizer for Neural Text Processing. In *EMNLP (Demonstration)*, January 2018.

LeCun, Y., Boser, B., Denker, J. S., Henderson, D., Howard, R. E., Hubbard, W., and Jackel, L. D. Backpropagation Applied to Handwritten Zip Code Recognition. *Neural Computation*, 1(4):541–551, December 1989. ISSN 0899-7667. doi: 10.1162/neco.1989.1.4.541.

Lee, J., Xiao, L., Schoenholz, S., Bahri, Y., Novak, R., Sohl-Dickstein, J., and Pennington, J. Wide Neural Networks of Any Depth Evolve as Linear Models Under Gradient Descent. In *Advances in Neural Information Processing Systems 32*, pp. 8570–8581. Curran Associates, Inc., 2019.

Martens, J. On the validity of kernel approximations for orthogonally-initialized neural networks. *arXiv:2104.05878 [cs]*, April 2021.

Martens, J., Ballard, A., Desjardins, G., Swirszcz, G., Dalibard, V., Sohl-Dickstein, J., and Schoenholz, S. S. Rapid training of deep neural networks without skip connections or normalization layers using Deep Kernel Shaping. *arXiv:2110.01765 [cs]*, October 2021.

Massaroli, S., Poli, M., Park, J., Yamashita, A., and Asama, H. Dissecting Neural ODEs. In *NeurIPS*, January 2020.

Merity, S., Xiong, C., Bradbury, J., and Socher, R. Pointer Sentinel Mixture Models. *arXiv:1609.07843 [cs]*, September 2016.

Mingo, J. A. and Speicher, R. *Free Probability and Random Matrices*. Fields Institute Monographs. Springer, New York, NY, 2017. ISBN 978-1-4939-6942-5. doi: 10.1007/978-1-4939-6942-5_1.

Nica, A. and Speicher, R. R-diagonal pairs - a common approach to Haar unitaries and circular elements. *arXiv:funct-an/9604012*, April 1996.

Pennington, J., Schoenholz, S., and Ganguli, S. Resurrecting the sigmoid in deep learning through dynamical isometry: Theory and practice. In *Advances in Neural Information Processing Systems*, volume 30. Curran Associates, Inc., 2017.

Poole, B., Lahiri, S., Raghu, M., Sohl-Dickstein, J., and Ganguli, S. Exponential expressivity in deep neural networks through transient chaos. In *Advances in Neural Information Processing Systems 29*, pp. 3360–3368. Curran Associates, Inc., 2016.

Revay, M., Wang, R., and Manchester, I. R. Lipschitz Bounded Equilibrium Networks, October 2020.

Saxe, A. M., McClelland, J. L., and Ganguli, S. Exact solutions to the nonlinear dynamics of learning in deep linear neural networks. *arXiv:1312.6120 [cond-mat, q-bio, stat]*, February 2014.

Schoenholz, S. S., Gilmer, J., Ganguli, S., and Sohl-Dickstein, J. Deep Information Propagation. *arXiv:1611.01232 [cs, stat]*, April 2017.

Tracy, C. A. and Widom, H. Level-spacing distributions and the Airy kernel. *Communications in Mathematical Physics*, 159(1):151–174, January 1994. ISSN 1432-0916. doi: 10.1007/BF02100489.

Winston, E. and Kolter, J. Z. Monotone operator equilibrium networks. *arXiv:2006.08591 [cs, stat]*, May 2021.

Xiao, L., Bahri, Y., Sohl-Dickstein, J., Schoenholz, S., and Pennington, J. Dynamical Isometry and a Mean Field Theory of CNNs: How to Train 10,000-Layer Vanilla Convolutional Neural Networks. In *Proceedings of the 35th International Conference on Machine Learning*, pp. 5393–5402. PMLR, July 2018.

Yang, G. Tensor Programs I: Wide Feedforward or Recurrent Neural Networks of Any Architecture are Gaussian Processes. *arXiv:1910.12478 [cond-mat, physics:math-ph]*, May 2021.

A. Linear DEQ theory

A.1. Basic definitions

In this section we derive some statistics relevant to the linear DEQ, defined implicitly by

$$\mathbf{z}^* = \mathbf{W}\mathbf{z}^* + \mathbf{x} \tag{32}$$

for an $N \times N$ matrix $\mathbf{W}$ and an $N \times 1$ dimensional vector $\mathbf{x}$. This equation has solution

$$\mathbf{z}^* = (\mathbf{I} - \mathbf{W})^{-1}\mathbf{x} \tag{33}$$

for $\mathbf{W}$ without eigenvalues equal to 1.

We will usually consider $\mathbf{W}$ to be drawn from a random matrix ensemble. The three ensembles we focus on will be the random Gaussian ensemble, the Gaussian orthogonal ensemble of random symmetric matrices, and the Haar-distributed orthogonal matrices. We will consider $\mathbf{W} = \sqrt{V}\mathbf{W}_0$, where $\mathbf{W}_0$ is drawn from these families and V is a fixed factor controlling the average squared singular value of $\mathbf{W}$.

We can think of $\mathbf{z}^*$ as the limit of the iterative map

$$\mathbf{z}_{t+1} = \mathbf{W}\mathbf{z}_t + \mathbf{x} \tag{34}$$

This map converges if and only if the eigenvalues of $\mathbf{W}$ have magnitude less than 1.

We will also be interested in the *untied weights* case where the iterative equation is given by

$$\mathbf{z}_{t+1} = \mathbf{W}_t\mathbf{z}_t + \mathbf{x} \tag{35}$$

where $\mathbf{W}_t$ are independent random matrices from the same ensemble. Here $\mathbf{z}_\infty = \lim_{t\to\infty} \mathbf{z}_t$ does not exist, but statistics like $||\mathbf{z}_\infty||^2 = \lim_{t\to\infty} \mathbf{z}_t \cdot \mathbf{z}_t$ do converge.

A.2. Infinite depth, infinite width limit

In (Feng & Kolter, 2020), the authors argue that DEQs under finite iteration meet the conditions for the Tensor Programs (TP) framework (Yang, 2021) to apply, allowing them to use the untied weights calculation to compute the finite-depth NTK. These results are then used to postulate a fixed point map for the infinite-depth DEQ.

However, the theory in (Yang, 2021) is valid only for programs of finite length. We show how this limitation can be overcome in the case of linear DEQs with random initialization. We will focus on the marginal distribution of the $\mathbf{z}_i^*$ for a single input $\mathbf{x}$; the covariance between pairs of inputs $\mathbf{x},\ \mathbf{y}$ follows naturally from our arguments.

The key idea is that $\mathbf{z}_t$ better and better approximates $\mathbf{z}^*$, as both N and t increase. Since $\mathbf{z}_t$ converges to a Gaussian in N, we can find better and better approximations to $\mathbf{z}^*$ which are more and more Gaussian (with statistics converging to the desired ones). By carefully increasing t and N together, we show convergence to the desired distribution.

It is helpful to prove a Lemma about the approximation scheme. Let $\mathbf{W}$ have i.i.d. Gaussian elements with variance V/N. We begin by computing $||\mathbf{z}^* - \mathbf{z}_t||_2^2/N$. We have:

$$\frac{1}{N}||\mathbf{z}^* - \mathbf{z}_t||_2^2 = \frac{1}{N}\mathbf{x}^{\mathrm{T}}\left(\sum_{k=t+1}^{\infty}\mathbf{W}^k\right)^{\mathrm{T}}\left(\sum_{k'=t+1}^{\infty}\mathbf{W}^{k'}\right)\mathbf{x} \tag{36}$$

Note that this power series representation is not guaranteed to exist for finite N; fluctuations may take the largest eigenvalue above 1 even when $V < 1$. However, the empirical spectrum of $\mathbf{W}$ converges in probability to the circular law with radius $\sqrt{V}$; therefore given any $\delta > 0$ there exists some N_1 such that the sum converges with probability at least $1 - \delta/2$ for all $N > N_1$.

Having established the convergence of the power series representation, factoring gives us

$$\frac{1}{N}||\mathbf{z}^* - \mathbf{z}_t||_2^2 = \frac{1}{N}(\mathbf{z}^*)^{\mathrm{T}}(\mathbf{W}^{t+1})^{\mathrm{T}}\mathbf{W}^{t+1}\mathbf{z}^* \tag{37}$$

As $N \to \infty$, the empirical distribution of $(\mathbf{W}^{t+1})^{\mathrm{T}}\mathbf{W}^{t+1}$ converges in probability to a distribution with a maximum eigenvalue bounded by tV^{t+1} (Pennington et al., 2017). Therefore, for every $\delta > 0$, there exists an $N_2 > N_1$ such that for all $N > N_2$, we have

$$\frac{1}{N}||\mathbf{z}^* - \mathbf{z}_t||_2^2 \leq \frac{1}{N}\mathbf{z}^* \cdot \mathbf{z}^*(t+1)V^{t+1} \tag{38}$$

with probability at least $1 - \delta/2$. This means that we have the (loose) bound

$$||(\mathbf{z}^*)_i - (\mathbf{z}_t)_i||_2^2 \leq (\mathbf{z}^* \cdot \mathbf{z}^*)(t+1)V^{t+1} \tag{39}$$

with probably at least $1 - \delta$.

We can use a similar argument to show that $\mathbf{z}^* \cdot \mathbf{z}^*$ converges in probability to $\frac{1}{1-V}\mathbf{x} \cdot \mathbf{x}$ (limiting value computed in Appendix C.2.2). We have the bound

$$|(\mathbf{z}^*)_i - (\mathbf{z}_t)_i|^2 < \frac{2t}{1-V}(\mathbf{x} \cdot \mathbf{x})V^{t+1} \tag{40}$$

where for any $\delta > 0$ the bound holds with probability at least $1 - \delta$ for all N with $N > N_3$ for some N_3.

We immediately see that for $V < 1$, the bound decreases with t, This gives us the following Lemma:

Lemma A.1. *Let $F_N(z)$ and $F_N(z;t)$ be the CDFs of $\mathbf{z}_i^*$ and $(\mathbf{z}_t)_i$. Fix z. Then for $\epsilon > 0$, there exist N_c and t_c such that*

$$|F_N(z) - F_N(z;t)| < \epsilon \tag{41}$$

for $t > t_c$, $N > N_c$.

Proof. Let $1 > \delta > 0$, $1 > \tilde{\epsilon} > 0$. There exists a t_c, N_c such that for $t > t_c$, $N > N_c$, $|(\mathbf{z}^*)_i - (\mathbf{z}_t)_i| < \tilde{\epsilon}$ for $t > t_c$, $N > N_c$ with probability at least $1 - \delta$ (for example, by choosing $t_c = O(\ln(V/\tilde{\epsilon}))$.

This means that we have

$$P(\mathbf{z}_i^* < z) \geq (1-\delta)P((\mathbf{z}_t)_i < z - \tilde{\epsilon}) \tag{42}$$

as well as

$$P(\mathbf{z}_i^* < z) \leq \frac{1}{1-\delta}P((\mathbf{z}_t)_i < z + \tilde{\epsilon}) \tag{43}$$

Note that there exists a constant B such that $F_N(z;t) - F_N(z';t) \leq B(z - z')$ for $z - z' < 1$. Therefore we have:

$$P(\mathbf{z}_i^* < z) \geq (1-\delta)[P((\mathbf{z}_t)_i < z) + B\tilde{\epsilon}] \tag{44}$$

$$P(\mathbf{z}_i^* < z) \leq \frac{1}{1-\delta}[P((\mathbf{z}_t)_i < z) + B\tilde{\epsilon}] \tag{45}$$

Therefore we have the bound:

$$|F_N(t,z) - F_N(t,z)| < 2(\delta + B\tilde{\epsilon}) \tag{46}$$

Given an $\epsilon > 0$, choose δ and $\tilde{\epsilon}$ such that $(\delta + B\tilde{\epsilon}) < \epsilon$. Then the lemma follows immediately from Equation 46 with t_c and N_c defined as above. □

Armed with this lemma, we can prove the following:

Theorem A.2. *Let $\mathbf{W}$ be i.i.d. Gaussian with elements of variance V/N, and let $\mathbf{z}_t$ and $\mathbf{z}^*$ be defined as above. Then for $V < 1$, as $N \to \infty$ each $\mathbf{z}_i^*$ converges in distribution to a Gaussian with mean $\mathbf{x}_i$ and second moment*

$$\mathrm{E}[(\mathbf{z}_i^*)^2] = \frac{1}{1-V}\frac{1}{N}\mathbf{x} \cdot \mathbf{x} \tag{47}$$

Proof. We will show that $\lim_{N\to\infty} F_N(z) \to F(z)$, where $F_N(z)$ are the CDFs of the $(\mathbf{z}_t)_i$ and $F(z)$ is the CDF of the Gaussian with appropriate parameters. We first note that the $(\mathbf{z}_t)_i$ converge in distribution to a Gaussian with mean $\mathbf{x}_i$ and second moment $m_2(t)$ given by

$$m_2(t) = \frac{1-V^t}{1-V}\frac{1}{N}\mathbf{x}\cdot\mathbf{x} \tag{48}$$

(as per (Feng & Kolter, 2020; Yang, 2021)). Let $F(z;t)$ be the corresponding limiting CDF. We note that

$$|\Phi(x/\sigma_1) - \Phi(x/\sigma_2)| \leq \left[\max_{x'} \frac{d\Phi(x'/\sigma)}{d\sigma}\bigg|_{\sigma=\sigma_1}\right](\sigma_1-\sigma_2) + O((\sigma_1-\sigma_2)^2) \tag{49}$$

where Φ is the standard Gaussian CDF. For fixed σ, the max derivative is bounded. Therefore we can write:

$$|F(z;t) - F(z)| \leq a(\sigma_t - \sigma_\infty) + b(\sigma_t - \sigma_\infty)^2 \tag{50}$$

for some constants a and b, where

$$\sigma_t^2 = \left(\frac{V^t}{1-V} - 1\right)\frac{1}{N}\mathbf{x}\cdot\mathbf{x} \tag{51}$$

We can re-write the bound as:

$$|F(z;t) - F(z)| \leq AV^t \tag{52}$$

for some fixed A.

Let $F_N(z;t)$ be the CDF of $(\mathbf{z}_t)_i$. We note that $\lim_{N\to\infty} F_N(z;t) = F(z;t)$. We will show that $F_N(z)$ can be made arbitrarily close to $F_N(z;t)$ for large N and t, allowing it to get arbitrarily close to $F(z)$. This will complete the proof.

Fix some z. Fix an $\epsilon > 0$. From Equation 52, there exists a t_b such that $|F(z;t) - F(z)| < \epsilon/3$ for all $t > t_b$. From Lemma A.1, there is a t_c and an N_c such that $|F_N(z;t) - F_N(z)| < \epsilon/3$.

Fix some t_a larger than t_b and t_c. Then there exists an N_a, greater than N_b and N_c, such that $|F_N(z;t_a) - F(z;t_a)| < \epsilon/3$ for all $N > N_a$. Using the triangle inequality, we have:

$$|F_N(z) - F(z)| < \epsilon \tag{53}$$

for all $N > N_a$. This completes the proof. □

Similar arguments apply to compute the covariate statistics of $\mathbf{z}^*(\mathbf{x})$, $\mathbf{z}^*(\mathbf{y})$ for different data points $\mathbf{x}$ and $\mathbf{y}$. This exact method of proof will not suffice for the non-linear case where the closed form solution is not known; however, in cases where the DEQ provably linearly converges to its fixed point, a similar argument may hold (depending on the properties of the Hessian).

A.3. Second moment

In the untied weights case, define

$$\mathrm{Var}[(\mathbf{z}_\infty)_i^2] \equiv \lim_{t\to\infty} \mathrm{Var}[(\mathbf{z}_t)_i^2] \tag{54}$$

For rotationally invariant $\mathbf{W}_t$ we have

$$\mathrm{Var}[(\mathbf{z}_\infty)_i^2] = \lim_{t\to\infty} \mathrm{Var}[(\mathbf{z}_i^*)^2] = \frac{1}{N}\left(\mathrm{E}[\mathbf{z}_t\cdot\mathbf{z}_t] - \mathrm{E}[\mathbf{z}_t]\cdot\mathrm{E}[\mathbf{z}_t]\right) \tag{55}$$

The latter gives $\mathbf{x}\cdot\mathbf{x}$. The first term can be computed recursively:

$$\mathrm{E}[\mathbf{z}_{t+1}\cdot\mathbf{z}_{t+1}] = \mathrm{E}\left[(\mathbf{W}_t\mathbf{z}_t + \mathbf{x})^{\mathrm{T}}(\mathbf{W}_t\mathbf{z}_t + \mathbf{x})\right] = \mathbf{x}\cdot\mathbf{x} + \mathrm{tr}[\mathbf{W}_t^{\mathrm{T}}\mathbf{W}_t]\mathrm{E}[\mathbf{z}_t\cdot\mathbf{z}_t] \tag{56}$$

Therefore we obtain the limit:

$$\mathrm{Var}[(\mathbf{z}_\infty)_i^2] = \frac{V}{1-V}\left(\frac{1}{N}\mathbf{x}\cdot\mathbf{x}\right) \tag{57}$$

valid for $V = \mathrm{tr}[\mathbf{W}_t^{\mathrm{T}}\mathbf{W}_t] < 1$.

Now consider the quantity $\mathrm{Var}[(\mathbf{z}_i^*)^2]$ in the tied weights case, averaged over the ensemble of $\mathbf{W}$. If $\mathbf{W}$ is from a rotationally invariant family, we have $\mathrm{Var}[(\mathbf{z}_i^*)^2] = \frac{1}{N}\left(\mathrm{E}[\mathbf{z}^* \cdot \mathbf{z}^*] - \mathrm{E}[\mathbf{z}^*] \cdot \mathrm{E}[\mathbf{z}^*]\right)$. For all rotationally invariant ensembles, we have $\mathrm{E}[\mathbf{z}^*] \cdot \mathrm{E}[\mathbf{z}^*] = \mathbf{x} \cdot \mathbf{x}$. The first term can be computed as: We have:

$$\mathrm{E}[\mathbf{z}^* \cdot \mathbf{z}^*] = \mathrm{E}\left[\mathbf{x}^{\mathrm{T}}(\mathbf{I} - \mathbf{W})^{-\mathrm{T}}(\mathbf{I} - \mathbf{W})^{-1}\mathbf{x}\right] = \mathrm{E}[\mathrm{tr}[(\mathbf{I} - \mathbf{W})^{-\mathrm{T}}(\mathbf{I} - \mathbf{W})^{-1}]]\mathbf{x} \cdot \mathbf{x} \tag{58}$$

If the spectral radius of $\mathbf{W}$ is less than 1, we can write the power series

$$\frac{1}{N}\mathrm{E}[\mathbf{z}^* \cdot \mathbf{z}^*] = \mathrm{E}\left[\mathrm{tr}\left[\sum_{j,k=0}^{\infty} (\mathbf{W}^j)^{\mathrm{T}}\mathbf{W}^k\right]\right] \tag{59}$$

For random and orthogonal $\mathbf{W}$, only terms with $j = k$ contribute and we have

$$\frac{1}{N}\mathrm{E}[\mathbf{z}^* \cdot \mathbf{z}^*] = \sum_{k=0}^{\infty} V^k = \frac{1}{1-V} \tag{60}$$

and we get $\mathrm{Var}[(\mathbf{z}^*)_i^2] = \frac{V}{1-V}$ as in the untied weights case.

However, for random symmetric matrices, all terms with $j + k$ even contribute. he semi-circular law gives us:

$$\mathrm{tr}[\mathbf{W}^{2k}] = C_k V^k,\ C_k = \frac{1}{k+1}\binom{2k}{k} \tag{61}$$

where C_k is the kth Catalan number. The trace vanishes for odd k. We also note that the Catalan numbers have the generating function $f_c(x)$ where

$$f_c(x) \equiv \sum_k C_k x^k = \frac{1 - \sqrt{1-4x}}{2x} \tag{62}$$

Computing, we have:

$$\mathrm{E}[(\mathbf{z}_i^*)^2] = \mathrm{tr}\left[\sum_{j,k=0}^{\infty} \mathbf{W}^j\mathbf{W}^k\right] = \sum_i (2i+1)V^i C_i \tag{63}$$

Using the theory of generating series, we have

$$\mathrm{E}[(\mathbf{z}_i^*)^2] = 2V f_c'(V) + f_c(V) \tag{64}$$

$$\mathrm{E}[(\mathbf{z}_i^*)^2] = 2V\left(-\frac{1-\sqrt{1-4V}}{2V^2} + \frac{1}{2V\sqrt{1-4V}}\right) + \frac{1-\sqrt{1-4V}}{2V} \tag{65}$$

$$\mathrm{E}[(\mathbf{z}_i^*)^2] = -\frac{1-\sqrt{1-4V}}{2V} + \frac{1}{\sqrt{1-4V}} \tag{66}$$

Therefore GOE distributed matrices have different statistics than the untied weights case, as well as the random and orthogonal DEQs. In particular, the diverging behavior is different. Here the divergence happens at $\sqrt{V} = \frac{1}{2}$ due to the radius of the semicircular law. In addition, if we write $\sqrt{V} = 1/2 - \delta$, for $\delta \ll 1$, the second moment is $O(\delta^{-1/2})$, compared to $O(\delta^{-1})$ for the other cases.

A.4. Fourth moment

The variability of DEQ outputs can be understood by computing the *length variance*. For $\mathbf{z} = \mathbf{M}\mathbf{x}$, with $\mathbf{x}$ i.i.d. Gaussian with 0 mean and variance $\sigma_{\mathbf{x}}^2$, we have

$$\sigma_{\mathbf{z}\cdot\mathbf{z}}^2 = \frac{1}{N}\mathrm{E}_{\mathbf{M}}[\mathrm{Var}_{\mathbf{x}}[\mathbf{z} \cdot \mathbf{z}]] \tag{67}$$

which by Lemma 2.2 is equivalent to

$$\sigma_{\mathbf{z}\cdot\mathbf{z}}^2 = \frac{1}{N}\mathrm{E}_{\mathbf{M}}[\mathrm{tr}((\mathbf{M}^{\mathrm{T}}\mathbf{M})^2)]\sigma_{\mathbf{x}}^4 \tag{68}$$

We give a proof of the lemma below.

Proof. Let $\{s\}$ be the singular values of $\mathbf{M}$, with associated singular vectors $\mathbf{v}_s$. We have:

$$\mathbf{z} \cdot \mathbf{z} = \sum_s s^2 |\mathbf{x} \cdot \mathbf{v}_s|^2 \tag{69}$$

Taking expectation over $\mathbf{x}$ we have

$$\mathrm{E}_{\mathbf{x}}[\mathbf{z} \cdot \mathbf{z}]^2 = \sigma_{\mathbf{x}}^4 \sum_{s,s'} s^2 (s')^2 \tag{70}$$

We also have

$$\begin{aligned}(\mathbf{z} \cdot \mathbf{z})^2 &= \left(\sum_s |s|^2 |\mathbf{x} \cdot \mathbf{v}_s|^2 \right)^2 \\ &= \sum_s |s|^4 |\mathbf{x} \cdot \mathbf{v}_s|^4 + \sum_{s \neq s'} |s'|^2 |s|^2 |\mathbf{x} \cdot \mathbf{v}_s|^2 |\mathbf{x} \cdot \mathbf{v}_{s'}|^2\end{aligned} \tag{71}$$

Note that $\mathrm{E}_{\mathbf{x}}[|\mathbf{x} \cdot \mathbf{v}_s|^4] = 3\sigma_{\mathbf{x}}^4$. This gives us:

$$\sigma_{\mathbf{z}\cdot\mathbf{z}}^2 = \frac{2}{N} \sigma_{\mathbf{x}}^4 \mathrm{E}_{\mathbf{M}} \left[\sum_s s^4 \right] = \frac{2}{N} \sigma_{\mathbf{x}}^4 \mathrm{E}_{\mathbf{M}} \left[\mathrm{tr}[(\mathbf{M}^{\mathrm{T}} \mathbf{M})^2] \right] \tag{72}$$

□

We will assume $\sigma_{\mathbf{x}}^2 = 1$ for the remainder of this section.

A.4.1. UNTIED WEIGHTS

In the untied weights case, we have

$$\sigma_\infty^2 = \frac{2}{N} \mathrm{E} \left[\mathrm{tr} \left[\sum_{j,k,l,m=0}^{\infty} \mathbf{M}_j^{\mathrm{T}} \mathbf{M}_k \mathbf{M}_l^{\mathrm{T}} \mathbf{M}_m \right] \right] \tag{73}$$

where $\mathbf{M}_k \equiv \prod_{t=1}^{k} \mathbf{W}_t$. The terms the terms that contribute are $j = k, l = m$ or $m = j, l = k$. If $j \neq l$, the two factors are freely independent and we can factor the trace. However, in the case $j = k = l = m$, the squared eigenvalues of $\mathbf{M}_j^{\mathrm{T}} \mathbf{M}_j$ contribute. This gives us:

$$\sigma_\infty^2 = \frac{2}{N} \sum_{i=0}^{\infty} 2(i+1)V^i - 2V^{2i} + B_i V^{2i} \tag{74}$$

where $B_i = \mathrm{E}[\mathrm{tr}[(\mathbf{M}_i^{\mathrm{T}} \mathbf{M}_i)^2]]$. This gives us:

$$\sigma_\infty^2 = \frac{2}{N} \left(\frac{2}{(1-V)^2} - \frac{2}{1-V^2} + \sum_{i=0}^{\infty} B_i V^{2i} \right) \tag{75}$$

In the random and GOE cases, as $N \to \infty$ $B_k = k + 1$. This gives us

$$\lim_{N\to\infty} \frac{N}{2} \sigma_\infty^2 = \frac{1}{(1-V^2)^2} + \frac{2}{(1-V)^2} - \frac{2}{1-V^2} \tag{76}$$

In the orthogonal case, $B_k = 1$. This gives us

$$\sigma_\infty^2 = \frac{2}{N} \left(\frac{2}{(1-V)^2} - \frac{1}{1-V^2} \right) \tag{77}$$

This is smaller than the GOE and random cases, but still has the same asymptotics as $V \to 1$.

Therefore in the untied weights case, σ_∞^2 scales as $(1-V)^{-2}$ in all cases. For the orthogonal case we have an analytic form for finite N, and for the random and GOE we have limiting behavior for $N \to \infty$. Evidently, as long as $V < 1$ we expect convergence of the statistics in the large N limit.

A.4.2. TIED WEIGHTS

Analogously to the untied weights case, we can attempt a power series solution:

$$\sigma^2_{\mathbf{z}^*\cdot\mathbf{z}^*} = \frac{2}{N}\mathrm{E}\left[\mathrm{tr}\left[\sum_{j,k,l,m=0}^{\infty}(\mathbf{W}^j)^{\mathrm{T}}(\mathbf{W}^k)(\mathbf{W}^l)^{\mathrm{T}}(\mathbf{W}^m)\right]\right] \tag{78}$$

This sum can be carried out in the orthogonal case. Only terms with $k+m=j+l$ contribute. Each side of the equation is independent, so for $k+m=i$ there are $(i+1)^2$ possible combinations. We have:

$$\sigma^2_{\mathbf{z}^*\cdot\mathbf{z}^*} = \frac{2}{N}\sum_{i=0}^{\infty}(i+1)^2V^i = \sum_{i=0}^{\infty}(i+1)(i+2)V^i - (i+1)V^i = \frac{2}{N}\left(\frac{2}{(1-V)^3} - \frac{1}{(1-V)^2}\right) \tag{79}$$

Here we see that the divergence is $(1-V)^{-3}$ - asymptotically different from the untied weights case.

For the random and GOE cases, the power series formulation diverges for finite N. This is due to the fluctuations in the largest eigenvalues. Even for $V<1$ for the random case ($V<\frac{1}{4}$ for the GOE case), there is non-zero probability that the spectral edge is greater than 1.

There are two ways to proceed. One is to compute $\sigma^2_{\mathbf{z}^*\cdot\mathbf{z}^*}$ using the definition involving $\mathrm{tr}[((\mathbf{I}-\mathbf{W})^{-\mathrm{T}}(\mathbf{I}-\mathbf{W})^{-1})^2]$. In the limit $N\to\infty$, this can be solved numerically using operator-valued free probability theory. The random and GOE cases are solved for this way in Appendix C.2, where we can obtain exact analytic solutions.

Another approach is to switch the order of limits. We define:

$$\sigma^2_{\mathbf{z}^*\cdot\mathbf{z}^*}(\infty) \equiv \lim_{L\to\infty}\lim_{N\to\infty}\mathrm{E}_{\mathbf{W}}\left[\mathrm{tr}\left[\sum_{j,k,l,m=0}^{L}(\mathbf{W}^j)^{\mathrm{T}}\mathbf{W}^k(\mathbf{W}^l)^{\mathrm{T}}\mathbf{W}^m\right]\right] \tag{80}$$

In the GOE case, we can write down an integral equation for $\sigma^2_{\mathbf{z}^*\cdot\mathbf{z}^*}(\infty)$:

$$\sigma^2_{\mathbf{z}^*\cdot\mathbf{z}^*}(\infty) = \frac{2}{\pi}\int_{-1}^{1}\frac{1}{(1-2\sqrt{V}x)^4}\sqrt{1-x^2}dx \tag{81}$$

There is a non-analytic formal power series solution. However, we can understand the behavior for V near the critical value analytically. Let $\sqrt{V}=1/2-\delta$, for some $\delta\ll 1$. Then, after shifting the integration variable we have:

$$\sigma^2_{\mathbf{z}^*\cdot\mathbf{z}^*}(\infty) = \frac{2}{\pi}\int_{0}^{2}\frac{1}{(2\delta+(1-2\delta)x)^4}\sqrt{x(2-x)}dx \tag{82}$$

In the limit of small δ, we have:

$$\sigma^2_{\mathbf{z}^*\cdot\mathbf{z}^*}(\infty) = \left(\frac{2}{\pi}\int_{0}^{2}\frac{1}{(2\delta+x)^4}\sqrt{x(2-x)}dx\right)(1+O(\delta)) \tag{83}$$

Let $y=x/2\delta$. Then we have:

$$\sigma^2_{\mathbf{z}^*\cdot\mathbf{z}^*}(\infty) \approx \frac{1}{2^{3/2}\pi}\delta^{-2.5}\int_{0}^{1/\delta}\frac{\sqrt{y(2-2\delta y)}}{(1+y)^4}dy \tag{84}$$

with relative error $O(\delta)$. As a further approximation, we have:

$$\sigma^2_{\mathbf{z}^*\cdot\mathbf{z}^*}(\infty) \approx \frac{1}{2\pi}\delta^{-2.5}\int_{0}^{\infty}\frac{\sqrt{y}}{(1+y)^4}dy \tag{85}$$

again with relative error $O(\delta)$. In total we have:

$$\sigma^2_{\mathbf{z}^*\cdot\mathbf{z}^*}(\infty) = \frac{1}{8}\delta^{-2.5}(1+O(\delta)) \tag{86}$$

for $\delta \ll 1$.

Since the squared second moment only scales as δ^{-2}, for small δ we have $\sigma^2_{\mathbf{z}^*\cdot\mathbf{z}^*}(\infty) = O(\delta^{-2.5})$.

This analysis lets us draw distinctions between the tied and untied cases, as well as between the matrix ensembles. In all cases, as expected the tied cases display more variance than the untied cases. These differences even show up asymptotically as δ, the distance to the transition, becomes small. The orthogonal case has one large advantage compared to the random and GOE cases: for any finite N, we expect convergence for V below the critical range. However, for both GOE and random matrices there is a chance of divergence for finite N, which increases as δ decreases.

However, in the intermediate regime where finite-N effects are small, the analysis of $\sigma^2_{\mathbf{z}^*\cdot\mathbf{z}^*}(\infty)$ suggests that the GOE *typically* has less variance than the orthogonal. The analytical results in Appendix C.2 and the numerical results in Figure 1 confirm this analysis.

B. Non-linear DEQ theory

B.1. DEQs in the infinite-width limit

Given a non-linear DEQ given by the implicit equation

$$\mathbf{z}^* = \phi(\mathbf{W}\mathbf{z}^*) + \mathbf{x} \tag{87}$$

which is sometimes thought of as an iterative equation of the form

$$\mathbf{z}_{t+1} = \phi(\mathbf{W}\mathbf{z}_t) + \mathbf{x} \tag{88}$$

we can analyze the properties of its infinite-width representations. The case of untied weights ($\mathbf{W}$ replaced by independent $\mathbf{W}_t$ in Equation 88) has been previously studied (Feng & Kolter, 2020).

In order to make infninite-width quantities well defined, we equip the DEQ layer with a readout vector $\mathbf{v}$ in order to construct the scalar function

$$f_{\mathbf{v},\mathbf{W}}(\mathbf{x}) \equiv \mathbf{v}^{\mathrm{T}}\mathbf{z}^*(\mathbf{x}) \tag{89}$$

We will omit the subscripts unless necessary.

The NNGP kernel (Lee et al., 2019) of f can be defined as:

$$\mathcal{K}(\mathbf{x},\mathbf{x}') = \mathrm{E}_{\mathbf{v},\mathbf{W}}[f(\mathbf{x})f(\mathbf{x}')] \tag{90}$$

which evaluates to:

$$\mathcal{K}(\mathbf{x},\mathbf{x}') = \frac{1}{N}\mathrm{E}_{\mathbf{v},\mathbf{W}}[\mathbf{z}^*(\mathbf{x})\cdot\mathbf{z}^*(\mathbf{x}')] \tag{91}$$

In order to compute the NTK, we must compute the derivative with respect to $\mathbf{W}$:

$$\frac{\partial f}{\partial \mathbf{W}} = \frac{\partial \mathbf{v}^{\mathrm{T}}\mathbf{z}^*}{\partial \mathbf{W}} \tag{92}$$

Using Equation 87 and the implicit function theorem, we have:

$$\frac{\partial f}{\partial \mathbf{W}} = (\mathbf{I} - \phi'(\mathbf{z}^*)\circ\mathbf{W})^{-\mathrm{T}}\mathbf{v}(\phi'(\mathbf{z}^*)\circ\mathbf{z}^*)^{\mathrm{T}} \tag{93}$$

For fixed $\mathbf{v}$, the NTK $\Theta(\mathbf{x},\mathbf{x}')$ with respect to $\mathbf{W}$ is given by:

$$\Theta(\mathbf{x},\mathbf{x}') \equiv \mathrm{E}\left[\frac{\partial f}{\partial \mathbf{W}}\cdot\frac{\partial f}{\partial \mathbf{W}}\right] \tag{94}$$

The NTK, after averaging over $\mathbf{v}$, is therefore:

$$\Theta(\mathbf{x},\mathbf{x}') = \mathrm{E}\left[\mathrm{tr}((\mathbf{I} - \phi'(\mathbf{z}^*)\circ\mathbf{W})^{-\mathrm{T}}(\mathbf{I} - \phi'((\mathbf{z}')^*)\circ\mathbf{W})^{-1})\,(\phi'(\mathbf{z}^*)\circ\mathbf{z}^*)\cdot(\phi'((\mathbf{z}')^*)\circ(\mathbf{z}')^*)\right] \tag{95}$$

The first term in Equation 95 gives the alignment of the Jacobians, computed using the implicit function theorem. The second term gives the alignment of the fixed points themselves, mediated by the derivative of the elementwise non-linearity ϕ. We expect (but do not prove here) that in many cases concentration of measure means the two terms are statistically independent.

We see from the form of the equation alone that, much like the linear case, the statistics of the matrix ensemble which $\mathbf{W}$ is drawn from affect the kernel, beyond simply the first and second moments. Orthogonal $\mathbf{W}$ should behave identically to random $\mathbf{W}$ in the large N limit; however other ensembles like the GOE will give us very different NTKs.

C. Free probability calculations

C.1. Spectrum of DW

Let $\mathbf{D}$ be a diagonal matrix, and $\mathbf{W}$ be a GOE matrix. If $\mathbf{D}$ is freely-independent of $\mathbf{W}$, and is non-negative, then we can use free probability theory to solve for the spectrum of $\mathbf{M} = \mathbf{DW}$. We first note that $\mathbf{DW}$ has the same spectrum as $\mathbf{D}^{1/2}\mathbf{W}\mathbf{D}^{1/2}$. This means that $\mathbf{DW}$ has real eigenvalues. We can use the Stieltjes transform of $\mathbf{DW}$ to solve for its spectrum using the theory of free multiplicative convolutions (Mingo & Speicher, 2017).

We review the basics of the theory here. We recall that the Stieltjes transform is given by:

$$G_{\mathbf{M}}(z) \equiv \mathrm{tr}\left[(z - \mathbf{M})^{-1}\right] \tag{96}$$

where tr is the normalized trace. The spectrum can be recovered from the Stieltjes transform via the relation

$$\rho(x) = -\frac{1}{\pi} \lim_{\epsilon \to 0^+} \Im[G_{\mathbf{M}}(x + i\epsilon)] \tag{97}$$

The Stieltjes transform is related to the moment generating function M by

$$M_{\mathbf{M}}(z) \equiv \sum_{k=1}^{\infty} m_k z^{-k} = zG_{\mathbf{M}}(z) - 1 \tag{98}$$

where $m_k = \mathrm{tr}[\mathbf{M}^k]$. Since $\mathbf{W}$ and $\mathbf{D}$ are freely independent with real spectra, we can compute the spectrum of their product using the S-transform - since the S-transform of a product of freely independent variables is the product of the S-transform. In terms of the moment generating function, we have:

$$S_{\mathbf{M}}(z) = \frac{1 + z}{zM^{-1}_{\mathbf{M}}(z)} \tag{99}$$

where M^{-1} is the functional inverse of the moment generating function. Given the S-transform, the MGF can be recovered using

$$M^{-1}_{\mathbf{M}}(z) = \frac{1 + z}{zS_{\mathbf{M}}(z)} \tag{100}$$

We begin by computing the S-transforms of $\mathbf{D}$ and $\mathbf{W}$. If $\mathbf{W}$ is a standard GOE element, we have:

$$G_{\mathbf{W}}(z) = \frac{z - \sqrt{z^2 - 4}}{2} \tag{101}$$

We use the branch of the square root function which has a branch cut on $[-2, 2]$, which corresponds to a cut on $(-\infty, 0]$ for the input of the square root. We choose the branch that has negative imaginary part just above the real axis.

This gives us a moment generating function of

$$M_{\mathbf{W}}(z) = \frac{z^2 - z\sqrt{z^2 - 4}}{2} - 1 \tag{102}$$

Let $w = M^{-1}_{\mathbf{W}}(z)$, the inverse MGF. We have:

$$2(z + 1) = w^2 - w\sqrt{w^2 - 4} \tag{103}$$

$$2(z+1) - w^2 = -w\sqrt{w^2-4} \tag{104}$$

$$4(z+1)^2 - 4(z+1)w^2 + w^4 = w^2(w^2-4) \tag{105}$$

$$4(z+1)^2 - 4zw^2 = 0 \tag{106}$$

This gives us an inverse moment generating function of

$$M_{\mathbf{W}}^{-1}(z) = \sqrt{\frac{(z+1)^2}{z}} \tag{107}$$

which can also be written as

$$M_{\mathbf{W}}^{-1}(z) = \sqrt{z + \frac{1}{z}} \tag{108}$$

This function has a branch cut from -1 to $+\infty$ (corresponding to $(-\infty, 0]$ for the argument of the square root), and takes on positive values for real negative z. This gives us the S-transform:

$$S_{\mathbf{W}}(z) = \frac{1+z}{z}\left[z + \frac{1}{z}\right]^{-1/2} \tag{109}$$

Therefore we have:

$$S_{\mathbf{M}}(z) = S_{\mathbf{W}}(z) S_{\mathbf{D}}(z) \tag{110}$$

When ϕ is the hart-tanh function, we can solve for the spectrum analytically. This means that the elements of $\mathbf{D}$ are Bernoulli random variables, with a probability $p(h^*)$ of being 1 which can be computed using the statistics of h^*. The Stieltjes transform is

$$G_{\mathbf{D}}(z) = \frac{1-p(h^*)}{z} + \frac{p(h^*)}{z-1} \tag{111}$$

The moment generating function is therefore

$$M_{\mathbf{D}}(z) = \sum_{k=1}^{\infty} p(h^*) z^{-k} = \frac{p(h^*)}{z}\left(1 - 1/z\right)^{-1} = \frac{p(h^*)}{z-1} \tag{112}$$

The inverse is given by

$$M_{\mathbf{D}}^{-1}(z) = \frac{p(h^*)}{z} + 1 = \frac{p(h^*)+z}{z} \tag{113}$$

The S-transform is given by

$$S_{\mathbf{D}}(z) = \frac{1+z}{z+p(h^*)} \tag{114}$$

Therefore, the overall S-transform is given by:

$$S_{\mathbf{M}}(z) = \frac{(1+z)^2}{z(z+p(h^*))}\left[z + \frac{1}{z}\right]^{-1/2} \tag{115}$$

The inverse moment generating function is given by

$$M_{\mathbf{M}}^{-1}(z) = \frac{z+p(h^*)}{z+1}\left[z + \frac{1}{z}\right]^{1/2} \tag{116}$$

which simplifies to

$$M_{\mathbf{M}}^{-1}(z) = \left(z + p(h^*)\right)[z]^{-1/2} \tag{117}$$

The moment generating function obeys the equation

$$z^2 M = (M + p(h^*))^2 \tag{118}$$

$$(M + p(h^*))^2 - z^2 M = 0 \tag{119}$$

$$M^2 + (2p(h^*) - z^2)M + p(h^*)^2 = 0 \tag{120}$$

Solving for the MGF, we have:

$$M_{\mathbf{M}}(z) = \frac{-(2p(h^*) - z^2) \pm \sqrt{(2p(h^*) - z^2)^2 - 4p(h^*)^2}}{2} \tag{121}$$

Simplification gives us:

$$M_{\mathbf{M}}(z) = -p(h^*) - \frac{z^2 \pm z\sqrt{z^2 - 4p(h^*)}}{2} \tag{122}$$

We expect the first moment to vanish and the second to be positive which gives us

$$M_{\mathbf{M}}(z) = -p(h^*) - \frac{z^2 - z\sqrt{z^2 - 4p(h^*)}}{2} \tag{123}$$

The Stieltjes transform is therefore given by

$$G_{\mathbf{M}}(z) = \frac{1 - p(h^*)}{z} - \frac{z - \sqrt{z^2 - 4p(h^*)}}{2} \tag{124}$$

Therefore, the spectrum of $\mathbf{M}$ is given by a combination of a delta function at 0 of weight $1 - p(h^*)$ and a semi-circular law with radius $2\sqrt{p(h^*)}$.

C.2. Operator valued free probability calculation of $\sigma^2_{\mathbf{z}^* \cdot \mathbf{z}^*}$

In this section we will compute the spectrum of $(\mathbf{I} - \mathbf{W})^{-\mathrm{T}}(\mathbf{I} - \mathbf{W})^{-1}$ for GOE and random $\mathbf{W}$ using operator valued free probability. In particular, we're interested in recovering the 2nd moment of the spectrum as V goes to 1.

The central object in operator-valued free probability theory is the *operator valued* Stieltjes transform. Let $\hat{\mathbf{P}}$ be a fixed $k \times k$ block matrix with $N \times N$ blocks. We define the function $\mathbf{G} : \mathbb{C}^{k \times k} \to \mathbb{C}^{k \times k}$ as

$$\mathbf{G}(\mathbf{B}) = \varphi[(\mathbf{B} - \hat{\mathbf{P}})^{-1}] \tag{125}$$

Here the φ operator applies the normalized trace to each $N \times N$ sub-block of the input.

As we will see, we can often compute the ordinary Stieltjes transform of rational function $f(\mathbf{W})$ by choosing some $\hat{\mathbf{P}}$ which is linear in $\mathbf{W}$ and $\mathbf{W}^{\mathrm{T}}$, such that for the appropriate choice of $\mathbf{B}$, $G_{f(\mathbf{W})}(z)$ is the first element of $\mathbf{G}(\mathbf{B})$. In this approach $\hat{\mathbf{P}}$ is often referred to as a *linear pencil*. We can then use well-developed techniques derived from free convolution theory to compute $\mathbf{G}(\mathbf{B})$. For a more detailed description of the techniques, we refer the reader to (Mingo & Speicher, 2017).

C.2.1. Warmup: semicircular case

For a GOE matrix $\mathbf{W}$, the matrix $(\mathbf{I} - \mathbf{W})^{-1}$ is already symmetric. Therefore its Stieltjes transform is well defined and analytic, and its fourth moment gives us the trace $\mathrm{tr}[((\mathbf{I} - \mathbf{W})^{-\mathrm{T}}(\mathbf{I} - \mathbf{W})^{-1})^2]$ needed to compute $\sigma^2_{\mathbf{z}^* \cdot \mathbf{z}^*}$.

Consider the block matrix $\mathbf{\Lambda}$ given by

$$\mathbf{\Lambda} = \begin{pmatrix} z & 0 \\ 0 & 0 \end{pmatrix} \tag{126}$$

where z is proportional to the $N \times N$ dimensional identity matrix $\mathbf{I}$. Then, if $\hat{\mathbf{P}}$ has the structure

$$\hat{\mathbf{P}} = \begin{pmatrix} 0 & \mathbf{V} \\ \mathbf{Q} & \mathbf{P} \end{pmatrix} \tag{127}$$

where 0 is the first $N \times N$ block, the operator valued Stieltjes transform has the following property:

$$\mathbf{G}(\mathbf{\Lambda})_{11} = \text{tr}[z + \mathbf{V}\mathbf{P}^{-1}\mathbf{Q}] = G_{-\mathbf{V}\mathbf{P}^{-1}\mathbf{Q}}(z) \tag{128}$$

With judicious choice of $\hat{\mathbf{P}}$, we can compute the Stieltjes transform of $(\mathbf{I} - \mathbf{W})^{-1}$. Consider

$$\hat{\mathbf{P}} = \begin{pmatrix} 0 & \mathbf{I} \\ \mathbf{I} & \mathbf{W} - \mathbf{I} \end{pmatrix} \tag{129}$$

Then, $\mathbf{G}(\mathbf{\Lambda})_{11} = G_{(\mathbf{I}-\mathbf{W})^{-1}}(z)$.

If $\hat{\mathbf{P}}$ can be decomposed into a sum $\hat{\mathbf{M}}$ + $\hat{\mathbf{F}}$, where $\hat{\mathbf{M}}$ is constant, and $\hat{\mathbf{F}}$ is a linear sum of (asymptotically) freely independent matrices with 0 trace, then we can solve for $\mathbf{G}(\mathbf{B})$. Define $\mathbf{Z} = \mathbf{B} - \hat{\mathbf{M}}$. Then we have:

$$\mathbf{Z}\mathbf{G} = \mathbf{I} + \eta(\mathbf{G})\mathbf{G} \tag{130}$$

where the covariance function $\eta : \mathbb{C}^{k\times k} \to \mathbb{C}^{k\times k}$ is defined as

$$\eta(\mathbf{G})_{ij} = \sum_{kl} \sigma(i,k;l,j)\mathbf{G}_{kl} \tag{131}$$

where

$$\sigma(i,k;l,j) \equiv \text{tr}[(\hat{\mathbf{F}})_{ik}(\hat{\mathbf{F}})_{lj}] \tag{132}$$

In this case, we have:

$$\eta(\mathbf{G}) = \begin{pmatrix} 0 & 0 \\ 0 & VG_{22} \end{pmatrix} \tag{133}$$

For $\mathbf{B} = \mathbf{\Lambda}$, we have

$$\mathbf{Z} = \begin{pmatrix} z & -\mathbf{I} \\ -\mathbf{I} & \mathbf{I} \end{pmatrix} \tag{134}$$

which gives the matrix equation:

$$\begin{pmatrix} zG_{11} - G_{21} & zG_{12} - G_{22} \\ -G_{11} + G_{21} & -G_{12} + G_{22} \end{pmatrix} = \begin{pmatrix} 1 & 0 \\ VG_{21}G_{22} & 1 + VG_{22}^2 \end{pmatrix} \tag{135}$$

We can now solve the system of equations for G_{11}.

We first note that, since $\mathbf{\Lambda}$ and $\hat{\mathbf{P}}$ are symmetric, so is $\mathbf{G}$. Therefore, $G_{12} = G_{21}$. The first equation gives us

$$zG_{11} - 1 = G_{21} = G_{12} \tag{136}$$

We recall that $zG_{11} - 1 = M$, where M is the moment generating function of $(\mathbf{I} - \mathbf{W})^{-1}$ (in z^{-1}). We will solve for M since the coefficient of the fourth order term gives us $\sigma^2_{\mathbf{z}^*\cdot\mathbf{z}^*}$. The second equation gives us

$$zG_{12} = zM = G_{22} \tag{137}$$

The fourth equation gives us

$$-M + zM = 1 + Vz^2M^2 \tag{138}$$

Solving for M, we have

$$M(z) = \frac{(z-1) \pm \sqrt{(z-1)^2 - 4Vz^2}}{2Vz^2} \tag{139}$$

We can choose the correct root by evaluating for small V. We know that

$$\lim_{V\to 0} M(z) = \frac{1}{z-1} \tag{140}$$

In particular this expansion is valid around large z. This means that the negative root is the correct one and we have

$$M(z) = \frac{(z-1) - \sqrt{(z-1)^2 - 4Vz^2}}{2Vz^2} \tag{141}$$

We can compute $\mathrm{tr}[((\mathbf{I}-\mathbf{W})^{-\mathrm{T}}(\mathbf{I}-\mathbf{W})^{-1})^2]$ by taking derivatives of $M(z)$. Let $w = z^{-1}$. Then we have:

$$\mathrm{tr}[((\mathbf{I}-\mathbf{W})^{-\mathrm{T}}(\mathbf{I}-\mathbf{W})^{-1})^2] = \frac{1}{4!}\left.\frac{d^4 M}{dw^4}\right|_{w=0} \tag{142}$$

Writing $M(w)$, we have

$$M(w) = w^2 \frac{(w^{-1}-1) - \sqrt{(1-w^{-1})^2 - 4Vw^{-2}}}{2V} \tag{143}$$

For ease of computation, we define

$$g(x) = \frac{-x - \sqrt{x^2 - 4V}}{2V} \tag{144}$$

Then we have:

$$M(w) = wg(w-1) \tag{145}$$

This gives us:

$$\frac{d^4 M}{dw^4} = 4\frac{d^3}{dw^3}g(w-1) + w\frac{d^4}{dw^4}g(w-1) \tag{146}$$

At $w = 0$, this evaluates to:

$$\left.\frac{d^4 M}{dw^4}\right|_{w=0} = \frac{2}{V}\left(\frac{3w}{(w^2-4V)^{3/2}} - \frac{3w^3}{(w^2-4V)^{5/2}}\right)\Bigg|_{w=-1} \tag{147}$$

which gives us

$$\mathrm{tr}[((\mathbf{I}-\mathbf{W})^{-\mathrm{T}}(\mathbf{I}-\mathbf{W})^{-1})^2] = \frac{1}{4V}\left(\frac{1}{(1-4V)^{5/2}} - \frac{1}{(1-4V)^{3/2}}\right) \tag{148}$$

for GOE distributed $\mathbf{W}$.

C.2.2. Random $\mathbf{W}$ case

The case of random $\mathbf{W}$ is more complicated, because the decomposition in Equation 130 depends on the constituent elements being semi-circular. We therefore first decompose $\mathbf{W}$ and $\mathbf{W}^{\mathrm{T}}$ into two freely-independent self-adjoint matrices:

$$\mathbf{X} \equiv \frac{\mathbf{W} + \mathbf{W}^{\mathrm{T}}}{2},\ \mathbf{Y} \equiv \frac{\mathbf{W} - \mathbf{W}^{\mathrm{T}}}{2i} \tag{149}$$

We can recover $\mathbf{W}$ via

$$\mathbf{X} + i\mathbf{Y} = \mathbf{W},\ \mathbf{X} - i\mathbf{Y} = \mathbf{W}^{\mathrm{T}} \tag{150}$$

With this decomposition, we can begin to take advantage of operator-valued free probability to write the Stieltjes transform of $(\mathbf{I}-\mathbf{W})^{-\mathrm{T}}(\mathbf{I}-\mathbf{W})^{-1}$ as part of a larger matrix.

We construct the linear pencil $\hat{\mathbf{P}}$ as follows. Cnsider the block matrix

$$\mathbf{P} = \begin{pmatrix} \mathbf{I} - 2\mathbf{X} & \mathbf{X} + i\mathbf{Y} \\ \mathbf{X} - i\mathbf{Y} & -\mathbf{I} \end{pmatrix} \tag{151}$$

Computing $\mathbf{P}^{-1}$, we have:

$$(\mathbf{P}^{-1})_{11} = (\mathbf{I} - 2\mathbf{X} + (\mathbf{X} + i\mathbf{Y})(\mathbf{X} - i\mathbf{Y}))^{-1} = (\mathbf{I} - 2\mathbf{X} + (\mathbf{X} + i\mathbf{Y})(\mathbf{X} - i\mathbf{Y}))^{-1} \tag{152}$$

$$(\mathbf{P}^{-1})_{11} = (\mathbf{I} - \mathbf{W} - \mathbf{W}^{\mathrm{T}} + \mathbf{W}\mathbf{W}^{\mathrm{T}})^{-1} = (\mathbf{I}-\mathbf{W})^{-\mathrm{T}}(\mathbf{I}-\mathbf{W})^{-1} \tag{153}$$

This gives us

$$\hat{\mathbf{P}} = \begin{pmatrix} 0 & \mathbf{I} & 0 \\ \mathbf{I} & -\mathbf{I} + 2\mathbf{X} & -\mathbf{X} - i\mathbf{Y} \\ 0 & -\mathbf{X} + i\mathbf{Y} & \mathbf{I} \end{pmatrix} \tag{154}$$

As desired, $\hat{\mathbf{P}}_{11}^{-1} = (\mathbf{I} - \mathbf{W})^{-\mathrm{T}}(\mathbf{I} - \mathbf{W})^{-1}$.

Repeating the setup of the operator-valued Stieltjes transform (this time for 3×3 complex matrices rather than 2×2, with the equivalent definition of $\mathbf{\Lambda}$, we have

$$\mathbf{Z} = \begin{pmatrix} z & -\mathbf{I} & 0 \\ -\mathbf{I} & \mathbf{I} & 0 \\ 0 & 0 & -\mathbf{I} \end{pmatrix} \tag{155}$$

for a complex z. We also define:

$$\tilde{\mathbf{X}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 2\mathbf{X} & -\mathbf{X} \\ 0 & -\mathbf{X} & 0 \end{pmatrix}, \; \tilde{\mathbf{Y}} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 0 & -i\mathbf{Y} \\ 0 & i\mathbf{Y} & 0 \end{pmatrix}, \tag{156}$$

So we have

$$\mathbf{\Lambda} - \hat{\mathbf{P}} = \mathbf{Z} - \tilde{\mathbf{X}} - \tilde{\mathbf{Y}} = \begin{pmatrix} z & -\mathbf{I} & 0 \\ -\mathbf{I} & \mathbf{I} - 2\mathbf{X} & \mathbf{X} + i\mathbf{Y} \\ 0 & \mathbf{X} - i\mathbf{Y} & -\mathbf{I} \end{pmatrix} \tag{157}$$

Equation 130 holds, where now the covariance map is defined by

$$\sigma(i, k; l, j) \equiv \mathrm{tr}[(\tilde{\mathbf{X}} + \tilde{\mathbf{Y}})_{ik}(\tilde{\mathbf{X}} + \tilde{\mathbf{Y}})_{lj}] \tag{158}$$

We can compute the individual covariance terms directly. Define $\boldsymbol{\sigma}_{ij} = \sigma(i, j; i, j)$. Then we have:

$$\boldsymbol{\sigma} = \begin{pmatrix} 0 & 0 & 0 \\ 0 & 2V & 0 \\ 0 & 0 & 0 \end{pmatrix} \tag{159}$$

We note that $\sigma(1j; lm) = \sigma(i1; lm) = \sigma(33; lm) = 0$ and

$$\sigma(22; 23) = \sigma(23; 22) = -V \tag{160}$$

Finally, $\sigma(23, 32) = \sigma(32, 23) = V$.

We can now write out a system of 9 equations that the Stieltjes transform obeys. We have:

$$\mathbf{ZG} = \begin{pmatrix} zG_{11} - G_{21} & zG_{12} - G_{22} & zG_{13} - G_{23} \\ -G_{11} + G_{21} & -G_{12} + G_{22} & -G_{13} + G_{23} \\ -G_{31} & -G_{32} & -G_{33} \end{pmatrix} \tag{161}$$

as well as

$$\eta(\mathbf{G}) = V \begin{pmatrix} 0 & 0 & 0 \\ 0 & 2G_{22} - (G_{23} + G_{32}) + G_{33} & -G_{22} \\ 0 & -G_{22} & G_{22} \end{pmatrix} \tag{162}$$

which gives us

$$(\eta(\mathbf{G})\mathbf{G})_{1j} = 0 \tag{163}$$

$$(\eta(\mathbf{G})\mathbf{G})_{2j} = V\left([2G_{22} - (G_{23} + G_{32}) + G_{33}]G_{2j} - G_{22}G_{3j}\right) \tag{164}$$

$$(\eta(\mathbf{G})\mathbf{G})_{3j} = VG_{22}(-G_{2j} + G_{3j}) \tag{165}$$

The top row of equations gives us:

$$zG_{11} - G_{21} = 1 \tag{166}$$

$$zG_{12} - G_{22} = 0 \tag{167}$$

$$zG_{13} - G_{23} = 0 \tag{168}$$

The middle row is:

$$-G_{11} + G_{21} = V\left([2G_{22} - (G_{23} + G_{32}) + G_{33}]G_{21} - G_{22}G_{31}\right) \tag{169}$$

$$-G_{12} + G_{22} = 1 + V\left([2G_{22} - (G_{23} + G_{32}) + G_{33}]G_{22} - G_{22}G_{32}\right) \tag{170}$$

$$-G_{13} + G_{23} = V\left([2G_{22} - (G_{23} + G_{32}) + G_{33}]G_{23} - G_{22}G_{33}\right) \tag{171}$$

The bottom row gives us:

$$-G_{31} = VG_{22}(-G_{21} + G_{31}) \tag{172}$$

$$-G_{32} = VG_{22}(-G_{22} + G_{32}) \tag{173}$$

$$-G_{33} = 1 + VG_{22}(-G_{23} + G_{33}) \tag{174}$$

We note that $\mathbf{\Lambda} - \hat{\mathbf{P}}$ is symmetric. Therefore $\mathbf{G}(\mathbf{\Lambda})$ is as well, and we can re-write the first set of equations as

$$zG_{11} - G_{12} = 1 \tag{175}$$

$$zG_{12} - G_{22} = 0 \tag{176}$$

$$zG_{13} - G_{23} = 0 \tag{177}$$

In the third set, the first two equations are now redundant. Therefore we have

$$-G_{23} = VG_{22}(-G_{22} + G_{23}) \tag{178}$$

$$-G_{33} = 1 + VG_{22}(-G_{23} + G_{33}) \tag{179}$$

From the second set, we have

$$-G_{12} + G_{22} = 1 + V\left([2G_{22} - 3G_{23} + G_{33}]G_{22}\right) \tag{180}$$

We can now attempt to solve for G_{11}. Using the first two equations, we have

$$G_{12} = zG_{11} - 1 \tag{181}$$

$$G_{22} = z(zG_{11} - 1) \tag{182}$$

Using the fifth equation, we have

$$G_{33} = \frac{VG_{22}G_{23} - 1}{1 + VG_{22}} \tag{183}$$

Substituting into the last equation, we have

$$-G_{12} + G_{22} = 1 + V\left(\left[2G_{22} - 3G_{23} + \frac{VG_{22}G_{23} - 1}{1 + VG_{22}}\right]G_{22}\right) \tag{184}$$

The fourth equation gives us

$$G_{23} = \frac{VG_{22}^2}{1 + VG_{22}} \tag{185}$$

Substitution gives us

$$-G_{12} + G_{22} = 1 + V\left(\left[2G_{22} - \frac{3VG_{22}^2}{1+VG_{22}} + \frac{V^2G_{22}^3}{(1+VG_{22})^2} - \frac{1}{1+VG_{22}}\right]G_{22}\right) \tag{186}$$

We define $M \equiv G_{12} = zG_{11} - 1$ (the moment generating function). Then we have:

$$-(1-z)M = 1 + V\left(\left[2zM - \frac{3V(zM)^2}{1+VzM} + \frac{V^2(zM)^3}{(1+VzM)^2} - \frac{1}{1+VzM}\right]zM\right) \tag{187}$$

We get the following cubic equation for M:

$$V^2z^2M^3 + 2VzM^2 + ((V-1)z + 1)M + 1 = 0 \tag{188}$$

The cubic can be solved for analytically using Cardano's formula to obtain the moment generating function. However, for the purposes of computing $\sigma^2_{\mathbf{z}^*\cdot\mathbf{z}^*}$ we simply need to compute m_2, where M has the expansion

$$M(z) = \sum_{k=1}^{\infty} m_k z^{-k} \tag{189}$$

The cubic in Equation 188 defines a second-order recursion for the m_k. We know that, by definition, there is no 0th order term in $M(z)$. The 0th order term in Equation 188 gives us $m_1 = \frac{1}{1-V}$ (as expected). The first order term (coefficient of z^{-1}) gives us

$$V^2m_1^3 + 2Vm_1^2 + (V-1)m_2 + m_1 = 0 \tag{190}$$

Solving for m_2, we have:

$$m_2 = \frac{V^2}{(1-V)^4} + \frac{2V}{(1-V)^3} + \frac{1}{(1-V)^2} \tag{191}$$

which matches up well with empirical measurements until $1 - V$ becomes small, when finite size effects begin to matter (Figure 1).

D. Experimental setup

We based our experiments on a Haiku implementation of a DEQ transformer (Khan, 2020) which uses the basic transformer layer (Al-Rfou' et al., 2019). We used a pre-trained sentencepiece tokenizer (Kudo & Richardson, 2018). We trained on Wikitext-103 (Merity et al., 2016) with a batch size of 512 and a context length of 128.

We added our own code to initialize the dense layer and the attention layers with different matrix families, though our experiments only modified the dense layers. Our hidden state had dimension 700. In the orthogonal and random cases, the dense layer expanded the state to dimension 2800 before projecting back down to 700. In the GOE case, since the matrices must be square, we kept the dimension at 700 throughout the network. We ran for 20 steps of the Broyden solver.

For the experiments with multiple seeds, we used a learning rate of 10^{-3} with a linear warmup for $2 \cdot 10^3$ steps, followed by a cosine learning rate decay for $5 \cdot 10^4$ steps. These parameters were chosen for their overall good performance on the random initialization with $\sqrt{V} = 0.1$. We found that increasing or decreasing the learning rate did not significantly improve performance in any case.