# The Emergence of Spectral Universality in Deep Networks: Supplementary Material

## 1 Review of free probability

For what follows, we define the key objects of free probability. Given a random matrix $\mathbf{X}$, its limiting spectral density is defined as

$$\rho_X(\lambda) \equiv \left\langle \frac{1}{N} \sum_{i=1}^{N} \delta(\lambda - \lambda_i) \right\rangle_X , \tag{S1}$$

where $\langle \cdot \rangle_X$ denotes an average w.r.t to the distribution over the random matrix $\mathbf{X}$. For large $N$, the empirical histogram of eigenvalues of a single realization of $\mathbf{X}$ converges to $\rho_X$. In turn, the *Stieltjes transform* of $\rho_X$ is defined as,

$$G_X(z) \equiv \int_{\mathbb{R}} \frac{\rho_X(t)}{z-t} dt \,, \qquad z \in \mathbb{C} \setminus \mathbb{R} \,, \tag{S2}$$

which can be inverted using,

$$\rho_X(\lambda) = -\frac{1}{\pi} \lim_{\epsilon \to 0^+} \operatorname{Im} G_X(\lambda + i\epsilon) \,. \tag{S3}$$

$G_X$ is related to the moment generating function $M_X$,

$$M_X(z) \equiv z G_X(z) - 1 = \sum_{k=1}^{\infty} \frac{m_k}{z^k} \,, \tag{S4}$$

where the $m_k$ is the $k$'th moment of the distribution $\rho_X$,

$$m_k = \int d\lambda \; \rho_X(\lambda) \lambda^k = \frac{1}{N} \langle \operatorname{tr} \mathbf{X}^k \rangle_X \,. \tag{S5}$$

In turn, we denote the functional inverse of $M_X$ by $M_X^{-1}$, which by definition satisfies $M_X(M_X^{-1}(z)) = M_X^{-1}(M_X(z)) = z$. Finally, the *S-transform* [9, 10] is defined in terms of the functional inverse $M_X^{-1}$ as,

$$S_X(z) = \frac{1+z}{z M_X^{-1}(z)} \,. \tag{S6}$$

The utility of the S-transform arises from its behavior under multiplication. Specifically, if $\mathbf{A}$ and $\mathbf{B}$ are two freely independent random matrices, then the S-transform of the product random matrix ensemble $\mathbf{AB}$ is simply the product of their S-transforms,

$$S_{AB}(z) = S_A(z) S_B(z) \,. \tag{S7}$$

## 2 Free probability and deep networks

We will now use eqn. (S7) to write down an implicit definition of the spectral density of $\mathbf{J}\mathbf{J}^T$, which is also the distribution of the square of the singular values of $\mathbf{J}$. Here $\mathbf{J}$ is the input-output Jacobian of a deep network defined in the main paper. First notice that, by eqn. (9), $M(z)$ and thus $S(z)$ depend only on the moments of the spectral density. The moments, in turn, can be defined in terms of traces (as in eqn. (S5)), which are invariant to cyclic permutations, i.e.,

$$\operatorname{tr}(\mathbf{A}^1 \mathbf{A}^2 \cdots \mathbf{A}^m)^k = \operatorname{tr}(\mathbf{A}^2 \cdots \mathbf{A}^m \mathbf{A}^1)^k \,. \tag{S8}$$

Therefore the S-transform is invariant to cyclic permutations. Now define matrices $\mathbf{Q}^l$ and $\tilde{\mathbf{Q}}^l$ as,

$$\begin{aligned} \mathbf{Q}^L &\equiv \mathbf{J}\mathbf{J}^T = (\mathbf{D}^L \mathbf{W}^L \mathbf{D}^{L-1} \cdots \mathbf{D}^1 \mathbf{W}^1)(\mathbf{D}^L \mathbf{W}^L \mathbf{D}^{L-1} \cdots \mathbf{D}^1 \mathbf{W}^1)^T \\ \tilde{\mathbf{Q}}^L &\equiv \left[ (\mathbf{D}^L \mathbf{W}^L)^T \mathbf{D}^L \mathbf{W}^L \right] (\mathbf{W}^{L-1} \mathbf{D}^{L-2} \cdots \mathbf{D}^1 \mathbf{W}^1)(\mathbf{W}^{L-1} \mathbf{D}^{L-2} \cdots \mathbf{D}^1 \mathbf{W}^1)^T \\ &= \left[ (\mathbf{D}^L \mathbf{W}^L)^T \mathbf{D}^L \mathbf{W}^L \right] \mathbf{Q}^{L-1} \,. \end{aligned} \tag{S9}$$

Now $\mathbf{Q}^L$ and $\tilde{\mathbf{Q}}^L$ are related by a cyclic permutation. Therefore the above argument shows that their S-transforms are equal, i.e. $S_{Q_L} = S_{\tilde{Q}_L}$. Furthermore $(\mathbf{D}^L\mathbf{W}^L)^T\mathbf{D}^L\mathbf{W}^L$ and $(\mathbf{D}^L)^2(\mathbf{W}^L)^T\mathbf{W}^L$ are related by a cyclic permutation, implying their S-transforms are also equal. Then a recursive application of eqn. (S7) and cyclic invariance of S-transforms implies that,

$$S_{JJ^T} = S_{Q_L} = S_{(D^L)^2} S_{(W^L)^T W^L} S_{Q^{L-1}} = \prod_{l=1}^{L} S_{(D^l)^2} S_{(W^l)^T W^l} = S_{D^2}^L S_{W^T W}^L \tag{S10}$$

where the last equality follows if each term in the Jacobian product identically distributed.

Given the expression for $S_{JJ^T}$, a simple procedure recovers the density of singular values of $\mathbf{J}$:

1. Use eqn. (S6) to obtain the moment generating function $M_{JJ^T}(z)$
2. Use eqn. (9) to obtain the Stieltjes transform $G_{JJ^T}(z)$
3. Use eqn. (S3) to obtain the spectral density $\rho_{JJ^T}(\lambda)$
4. Use the relation $\lambda = \sigma^2$ to obtain the density of singular values of $J$.

So in order to compute the distribution of singular values of of $J$, all that remains is to compute the S-transforms of $W^TW$ and of $D^2$. We will attack this problem for specific activation functions and matrix ensembles in the following sections.

## 3 Derivation of master equations for the spectrum of the Jacobian

To derive the master equation, we first insert (S6), for $\mathbf{X} = \mathbf{D^2}$, into (S10) to obtain

$$S_{JJ^T} = S_{W^TW}^L \left(\frac{1+z}{z}\right)^L \left(M_{D^2}^{-1}\right)^{-L} .$$

Then we find $M_{JJ^T}^{-1} = (1+z)(zS_{J^TJ})^{-1}$ by inverting (S6), which combined with the above equation yields

$$M_{JJ^T}^{-1} = S_{W^TW}^{-L} \left(\frac{1+z}{z}\right)^{1-L} \left(M_{D^2}^{-1}\right)^L .$$

Then solving for $M_{D^2}^{-1}$ yields

$$M_{D^2}^{-1} = \left(M_{JJ^T}^{-1} S_{W^TW}^L \left(\frac{1+z}{z}\right)^{L-1}\right)^{\frac{1}{L}} .$$

Applying $M_{D^2}$ to both sides gives,

$$z = M_{D^2}\left(\left(M_{JJ^T}^{-1} S_{W^TW}^L(z) \left(\frac{1+z}{z}\right)^{L-1}\right)^{\frac{1}{L}}\right) .$$

Finally, evaluating this equation at $z = M_{JJ^T}$ gives our sought after master equation:

$$M_{JJ^T}(z) = M_{D^2}\left(z^{\frac{1}{L}} S_{W^TW}\big(M_{JJ^T}(z)\big) \left(1 + \frac{1}{M_{JJ^T}(z)}\right)^{1-\frac{1}{L}}\right) . \tag{S11}$$

This is an implicit functional equation for $M_{JJ^T}(z)$, an unknown quantity, in terms of the known functions $M_{D^2}(z)$ and $S_{W^TW}(z)$. Furthermore, by substituting (S4), $M_{JJ^T} = zG_{JJ^T} - 1$, into (S11), we also obtain an implicit functional equation for the Stieltjes transform $G$ of $\rho_{JJ^T}(\lambda)$,

$$zG - 1 = M_{D^2}\left(z^{\frac{1}{L}} S_{W^TW}(zG-1) \left(\frac{zG}{zG-1}\right)^{1-\frac{1}{L}}\right). \tag{S12}$$

## 4 Derivation of Moments of deep spectra

The moments $m_k$ of the spectrum of $\mathbf{J}\mathbf{J}^T$ are encoded in the moment generating function

$$M_{JJ^T}(z) \equiv \sum_{k=1}^{\infty} \frac{m_k}{z^k} \,. \tag{S13}$$

These moments in turn can be computed in terms of the series expansions of $S_{W^TW}$ and $M_{D^2}$, which we define as

$$S_{W^TW}(z) \equiv \quad \sigma_w^{-2} \left(1 + \textstyle\sum_{k=1}^{\infty} s_k z^k\right) \tag{S14}$$

$$M_{D^2}(z) \equiv \qquad \textstyle\sum_{k=1}^{\infty} \frac{\mu_k}{z^k} \,, \tag{S15}$$

where the moments $\mu_k$ of $\mathbf{D}^2$ are given by,

$$\mu_k = \int \mathcal{D}h \; \phi'(\sqrt{q^*}h)^{2k} \,. \tag{S16}$$

We can substitute these moment expansions into (S11) to obtain equations for the unknown moments $m_k$ of the spectrum of $\mathbf{J}\mathbf{J}^T$, in terms of the known moments $\mu_k$ and $s_k$. We can solve for the low order moments by expanding (S11) in powers of $z^{-1}$. By equating the coefficients of $z^{-1}$ and $z^{-2}$, we obtain the following equations for $m_1$ and $m_2$,

$$\begin{aligned} m_1 &= \sigma_w^2 \mu_1 m_1^{1-\frac{1}{L}} \\ m_2 &= \sigma_w^4 \mu_2 m_1^{2-\frac{2}{L}} \\ &\quad + \sigma_w^2 \mu_1 m_1^{2-\frac{1}{L}} \left( \left( \frac{m_2}{m_1^2} - 1 \right) \left( 1 - \frac{1}{L} \right) - s_1 \right) . \end{aligned} \tag{S17}$$

Solving for $m_1$ and $m_2$ yields,

$$\begin{aligned} m_1 &= (\sigma_w^2 \mu_1)^L \\ m_2 &= (\sigma_w^2 \mu_1)^{2L} \, L \left( \frac{\mu_2}{\mu_1^2} + \frac{1}{L} - 1 - s_1 \right) . \end{aligned} \tag{S18}$$

## 5 Transforms of Nonlinearities

Here we compute the moment generating functions $M_{D^2}(z)$ for various choices of the nonlinearity $\phi$, some of which are displayed in Table 1 of the main paper.

### 5.1 $\phi(x) = x$

$$\begin{aligned} M_{D^2}(z) &= \int \mathcal{D}x \frac{1}{z-1} \\ &= \frac{1}{z-1} \,. \end{aligned} \tag{S19}$$

### 5.2 $\phi(x) = [x]_+$

$$\begin{aligned} M_{D^2}(z) &= \int \mathcal{D}x \frac{\theta(x)^2}{z - \theta(x)^2} \\ &= \frac{1}{2} \int \mathcal{D}x \frac{1}{z-1} \\ &= \frac{1}{2} \frac{1}{z-1} \,. \end{aligned} \tag{S20}$$

**5.3** $\phi(x) = \text{htanh}(x)$

$$\begin{aligned} M_{D^2}(z) &= \int \mathcal{D}x \frac{\theta(1-q_* x)^2}{z - \theta(1+q_* x)^2} \\ &= \text{erf}(\frac{1}{\sqrt{2}q_*}) \int \mathcal{D}x \frac{1}{z-1} \\ &= \text{erf}(\frac{1}{\sqrt{2}q_*}) \frac{1}{z-1} \,. \end{aligned} \tag{S21}$$

**5.4** $\phi(x) = [x]_+ + \alpha[-x]_+$

$$\begin{aligned} M_{D^2}(z) &= \int \mathcal{D}x \frac{\phi'(q_* x)^2}{z - \phi'(q_* x)^2} \\ &= \frac{1}{2(z-1)} + \frac{1}{2(z/\alpha^2 - 1)} \,. \end{aligned} \tag{S22}$$

**5.5** $\phi(x) = \text{erf}(\frac{\sqrt{\pi}}{2}x)$

$$\begin{aligned} M_{D^2}(z) &= \int \mathcal{D}x \frac{\phi'(q_* x)^2}{z - \phi'(q_* x)^2} \\ &= \sum_{k=1}^{\infty} z^{-k} \int \mathcal{D}x \; e^{-\frac{1}{2}\pi k q_*^2 x^2} \\ &= \sum_{k=1}^{\infty} \frac{1}{z^k \sqrt{1 + \pi k q_*^2}} \\ &= \frac{1}{\sqrt{\pi} q_* z} \Phi\left(\frac{1}{z}, \frac{1}{2}, 1 + \frac{1}{\pi q_*^2}\right) \,, \end{aligned} \tag{S23}$$

where $\Phi$ is the special function known as the Lerch transcendent.

**5.6** $\phi(x) = \frac{2}{\pi}\arctan(\frac{\pi}{2}x)$

$$\begin{aligned} M_{D^2}(z) &= \int \mathcal{D}x \frac{\phi'(q_* x)^2}{z - \phi'(q_* x)^2} \\ &= \int \mathcal{D}x \; \frac{16}{(4 + \pi^2 q_*^2 x^2)^2 z + 16} \\ &= -\frac{\sqrt{2}}{\pi^{3/2} q_*^2 \sqrt{z}} \left( \frac{e^{\frac{z_+}{2}}}{\sqrt{z_+}} \,\text{erfc}\left(\sqrt{\frac{z_+}{2}}\right) - \frac{e^{\frac{z_-}{2}}}{\sqrt{z_-}} \,\text{erfc}\left(\sqrt{\frac{z_-}{2}}\right) \right) \,, \end{aligned} \tag{S24}$$

where,

$$z_\pm = \frac{4(\sqrt{z} \pm 1)}{\pi^2 q_*^2 \sqrt{z}} \,. \tag{S25}$$

## 6 Transforms of Weights

First consider the case of an orthogonal random matrix satisfying $\mathbf{W^T W} = \mathbf{I}$. Then

$$\begin{aligned} \rho_{W^T W}(\lambda) &= \delta(\lambda - 1) \\ G_{W^T W}(z) &= (z-1)^{-1} \\ M_{W^T W}(z) &= (z-1)^{-1} \\ M^{-1}_{W^T W}(z) &= (1+z)/z \\ S_{W^T W}(z) &= 1. \end{aligned} \tag{S26}$$

The case of a random Gaussian random matrix $\mathbf{W}$ with zero mean, variance $\frac{1}{N}$ entries is more complex, but well known:

$$\begin{aligned}
\rho_{W^TW}(\lambda) &= (2\pi)^{-1}\sqrt{4-\lambda} \quad \text{for } \lambda \in [0,4] \\
G_{W^TW}(z) &= \frac{1}{2}\left(1 - \sqrt{\frac{z-4}{2}}\right) \\
M_{W^TW}(z) &= \frac{1}{2}(z - \sqrt{z(z-4)} - 2) \\
M^{-1}_{W^TW}(z) &= (1+z)^2/z \\
S_{W^TW}(z) &= (1+z)^{-1}.
\end{aligned} \tag{S27}$$

Furthermore, by scaling $\mathbf{W} \to \sigma_w \mathbf{W}$, the S-transform scales as $S_{W^TW} \to \sigma_w^{-2} S_{W^TW}$, yielding the S-transforms in Table 1.

Table 1: Transforms of weights

| Random Matrix $\mathbf{W}$ | $S_{W^TW}(z)$ | $s_1$ |
|---|---|---|
| Scaled Orthogonal | $\sigma_w^{-2}$ | 0 |
| Scaled Gaussian | $\sigma_w^{-2}(1+z)^{-1}$ | $-1$ |

## 7 Universality class of orthogonal Hard Tanh networks

We consider hard tanh with orthogonal weights. The moment generating function is,

$$M_{D^2} = \mathrm{erf}(\frac{1}{\sqrt{2q^*}})\frac{1}{z-1}\,, \tag{S28}$$

so that

$$\frac{\mu_2}{\mu_1^2} = \frac{1}{\mathrm{erf}(\frac{1}{\sqrt{2q^*}})} \quad \text{and} \quad g = \frac{1}{\sqrt{\mu_1}} = \frac{1}{\mathrm{erf}(\frac{1}{\sqrt{2q^*}})}\,. \tag{S29}$$

Also we have,

$$\sigma^2_{JJ^T} = L\left(\frac{1}{\mathrm{erf}(\frac{1}{\sqrt{2q^*}})} - 1\right) \quad \Rightarrow \quad q_*(L) = \frac{1}{\sqrt{2\,\mathrm{erf}^{-1}(\frac{L}{L+\sigma_0^2})}}\,. \tag{S30}$$

if we wish to scale $q^*$ with depth $L$ so as to achieve a depth independent constant variance $\sigma^2_{JJ^T} = \sigma_0^2$ as $L \to \infty$. This expression for $q^*$ gives,

$$M_{D^2} = \frac{L}{L+\sigma_0^2}\frac{1}{z-1} \quad \text{and} \quad \mu_1 = \frac{L}{L+\sigma_0^2}\,, \tag{S31}$$

so that,

$$S_{JJ^T} = S_{D^2} = (\mu_1 \frac{1+z}{z M^{-1}_{D^2}})^L = \left(\frac{L(1+z)}{L(1+z)+z\sigma_0^2}\right)^L = \left(1 + \frac{z\sigma_0^2}{L(1+z)}\right)^{-L}\,. \tag{S32}$$

The large depth limit gives,

$$S_{JJ^T} = e^{-\frac{z\sigma_0^2}{(1+z)}}\,. \tag{S33}$$

Solving for $G(z)$ gives,

$$G(z) = \frac{1}{z}\frac{1}{1 + W(\frac{-\sigma_0}{z})/\sigma_0^2}\,, \tag{S34}$$

where $W$ is the standard Lambert-W function, or product log. The derivative of this function has double poles at,

$$\lambda_0 = 0, \quad \lambda_2 = e^{\sigma_0^2}\,, \tag{S35}$$

which are locations where the spectral density diverges. There is also a single pole at,

$$\lambda_1 = \sigma_0^2 e\,, \tag{S36}$$

which is the maximum value of the bulk of the density.

## 8 Universality class of orthogonal erf networks

Consider $\phi(x) = \sqrt{\frac{\pi}{2}}\,\mathrm{erf}(\frac{x}{\sqrt{2}})$, which has been scaled so that $\phi'(0) = 1$ and $\phi'''(0) = -1$. The $\mu_k$ are given by,

$$\mu_k = \frac{1}{\sqrt{1+2kq^*}} \tag{S37}$$

so that

$$\sigma^2_{JJ^T} = L\left(\frac{1+2q^*}{\sqrt{1+4q^*}} - 1\right). \tag{S38}$$

If we wish to scale $q^*$ with depth $L$ so as to achieve a depth independent constant variance $\sigma^2_{JJ^T} = \sigma_0^2$ as $L \to \infty$, then we can choose

$$q^*(L) = \left(\frac{\sigma_0^2}{L} + \frac{\sigma_0^4}{2L^2} + \left(\frac{\sigma_0}{2L} + \frac{\sigma_0^3}{2L^2}\right)\sqrt{2L+\sigma_0^2}\right)^{1/4}. \tag{S39}$$

Since we also assume the network is critical, we also have that,

$$\sigma_w^2 = (1+2q^*)^{\frac{1}{4}}\,. \tag{S40}$$

To illustrate universality, we next consider an arbitrary activation function, and assume that it has a Taylor expansion around 0. This allows us to expand the $\mu_k$. First we write,

$$\phi(x) = \sum_{k=0}^{\infty} \phi_k x^k\,, \tag{S41}$$

We will need $\phi_1 \neq 0$. First we will assume that $\phi_2 \neq 0$. Using this expansion we can write,

$$\mu_k = \phi_1^{2k}\left(1 + k\Big((2k-1)\frac{\phi_2^2}{\phi_1^2} + \frac{\phi_3}{\phi_1}\Big)q_*^2 + \mathcal{O}(q_*^4)\right). \tag{S42}$$

We also have

$$S_{JJ^T} = \left(\mu_1 \frac{1+z}{zM^{-1}_{D^2}}\right)^L, \tag{S43}$$

where we have used the fact that the network is critical so that we have $\mu_1 = g^{-2}$. Using the Lagrange inversion theorem to expand $M^{-1}_{D^2}$, we find that

$$\mu_1 \frac{1+z}{zM^{-1}_{D^2}} = 1 - 4\frac{\phi_2^2}{\phi_1^2} z q_*^2 + \mathcal{O}(q_*^4)\,. \tag{S44}$$

Meanwhile,

$$\begin{aligned} L &= \sigma_0^2\left(\frac{\mu_2}{\mu_1^2} - 1\right)^{-1}, \\ &= \sigma_0^2 \frac{\phi_1^2}{4\phi_2^2 q_*^2}\,, \end{aligned} \tag{S45}$$

so that,

$$\begin{aligned} S_{JJ^T} &= \left(\mu_1 \frac{1+z}{z M_{D^2}^{-1}}\right)^L \\ &= \left(1 - \frac{\sigma_0^2}{L} z\right)^L \\ &= e^{-\sigma_0^2 z} + \mathcal{O}(L^{-1}) \,, \end{aligned} \tag{S46}$$

Next we will assume that $\phi_2 = 0$ and $\phi_3 \neq 0$[1]. Using the above expansion we can write,

$$\mu_k = \phi_1^{2k} \left(1 + k \frac{\phi_3}{\phi_1} q_*^2 + \mathcal{O}(q_*^4)\right) . \tag{S47}$$

Also we have,

$$S_{JJ^T} = \left(\mu_1 \frac{1+z}{z M_{D^2}^{-1}}\right)^L , \tag{S48}$$

where we have used the fact that the network is critical so that we have $\mu_1 = \sigma_w^{-2}$. Using the Lagrange inversion theorem to expand $M_{D^2}^{-1}$, we find that

$$\mu_1 \frac{1+z}{z M_{D^2}^{-1}} = 1 - 2 \frac{\phi_3^2}{\phi_1^2} z q_*^4 + \mathcal{O}(q_*^6) \,. \tag{S49}$$

Meanwhile,

$$\begin{aligned} L &= \sigma^2 \left(\frac{\mu_2}{\mu_1^2} - 1\right)^{-1} , \\ &= \sigma^2 \frac{\phi_1^2}{2\phi_3^2 q_*^4} \,, \end{aligned} \tag{S50}$$

so that,

$$\begin{aligned} S_{JJ^T} &= \left(\mu_1 \frac{1+z}{z M_{D^2}^{-1}}\right)^L \\ &= \left(1 - \frac{\sigma_0^2}{L} z\right)^L \\ &= e^{-\sigma_0^2 z} + \mathcal{O}(L^{-1}) \,, \end{aligned} \tag{S51}$$

establishing a universal limiting S-transform (subject to our assumptions). From this result we can extract the Stieltjes transform and thus the spectral density. The result establishes a universal double scaling limiting spectral distribution.

Next we observe that the Stieltjes transform can be expressed in terms of a generalization of the Lambert -$W$ function called the r-Lambert function, $W_r(z)$, which is defined by

$$W_r e^{W_r} + r W_r = z \,. \tag{S52}$$

In terms of this function, the Stieltjes transform is,

$$G(z) = \frac{W_{-e^{\sigma_0^2} z}(-\sigma_0^2 z e^{\sigma_0^2})}{z \sigma_0^2} \,. \tag{S53}$$

[1] We suspect these additional assumptions are unnecessary and that the results which follow are valid so long as there exists a $k$ for which $\phi_k \neq 0$. It would be interesting to prove this.

We can extract the maximum and minumum eigenvalue by finding the branch points of this function. It suffices to look for poles in the derivative of the numerator of $G(z)$. Using $r = -\sigma_0^2 z e^{\sigma_0^2}$, eqn. (S52) and its total derivative with respect to $z$ yields the following equation defining the locations of these poles,

$$e^{W_r}\big(1 + W_r\big) = z e^{z^2}\,, \tag{S54}$$

which is solved by

$$W_r = W(e^{1+\sigma_0^2} z) - 1\,, \tag{S55}$$

where $W$ is the standard Lambert W function. Next we substite this relation into eqn. (S52); zeros in $z$ then define the location of the branch points. Some straightforward algebra yields the maximum and minimum eigenvalue,

$$\lambda_\pm = \frac{1}{2} e^{-\frac{1}{2}\sigma_\pm^2}\left(2 + \sigma_\mp^2\right)\,, \quad \text{where} \quad \sigma_\pm^2 = \sigma_0\left(\sigma_0 \pm \sqrt{\sigma_0^2 + 4}\right) \tag{S56}$$

## 9 Orthogonal weights are required for stable, universal limiting distributions

We work at criticality so $\chi = \sigma_w^2 \mu_1 = 1$. This implies that

$$\begin{aligned} \mu_{JJ^T} &= m_1 = 1 \\ \sigma_{JJ^T}^2 &= m_2 - m_1^2 = L\left(\frac{\mu_2}{\mu_1^2} - 1 - s_1\right)\,. \end{aligned} \tag{S57}$$

Observe that Jensen's inequality requires that $\mu_2 \geq \mu_1^2$. If we require that $\sigma_{JJ^T}^2$ approach a constant as $L \to \infty$, we must have that,

$$s_1 \geq 0\,. \tag{S58}$$

Similarly, writing

$$M_{W^TW}(z) = \sum_{k=1}^{\infty} \frac{\mathfrak{m}_k}{z^k}\,, \tag{S59}$$

we can relate $\sigma_w$ and $s_1$ to $\mathfrak{m}_1$ and $\mathfrak{m}_2$. Specifically, evaluating the relation,

$$M^{-1}_{W^TW}(z) = \frac{1+z}{z S_{W^TW}(z)}\,, \tag{S60}$$

at $z = M_{W^TW}(x)$, gives,

$$x = \frac{1 + M_{W^TW}(x)}{M_{W^TW}(x) S_{W^TW}(M_{W^TW}(x))}\,. \tag{S61}$$

Expanding this equation to second order gives,

$$\begin{aligned} \mathfrak{m}_1 &= g^2 \\ \mathfrak{m}_2 &= g^2 \mathfrak{m}_1 (1 - s_1)\,. \end{aligned} \tag{S62}$$

Finally we see that,

$$\sigma_{WW^T}^2 = \mathfrak{m}_2 - \mathfrak{m}_1^2 = -g^4 s_1\,. \tag{S63}$$

Positivity of variance gives $s_1 \leq 0$, which, together with eqn. (S58) implies,

$$s_1 = 0\,. \tag{S64}$$

Altogether we see that the variance of the distribution of eigenvalues of $WW^T$ must be zero. Since its mean is equal to $\sigma_w^2$, we see that the only valid distribution for the eigenvalues of $WW^T$ is a delta function peaked at $\sigma_w^2$, i.e. the distribution corresponding to the singular values of an orthogonal matrix scaled by $\sigma_w$.