Problem Set 4: Jacobian spectra calculations

Depth First Learning week 5

Problem 1: Set up

In this problem set, we perform the main calculations from the *Resurrecting the Sigmoid* paper. The ultimate aim is to look for conditions under which we can achieve *dynamical isometry*, the condition that all of the singular values of the network's Jacobian have magnitude 1. Thus, the problems in this set are all aimed at calculating the eigenvalue spectral density $\rho_{JJ^T}(\lambda)$ of nets' Jacobians for specific choices of nonlinearities and weight-matrix initializations. We accomplish this by using the rule we learned from free probability: S-transforms of freely-independent matrix ensembles multiply under matrix multiplication. Following this logic, we will calculate S-transforms for the matrices WW^T and D^2, combine these results to arrive at S_{JJ^T}, and from that calculate $\rho_{JJ^T}(\lambda)$. In this problem set, as in the paper, we do not prove that the matrices are freely independent, but instead take that as an assumption.

Recall that our neural network is defined by the relations:

$$h^l = W^l x^{l-1} + b^l \tag{1}$$
$$x^l = \phi(h^l), \tag{2}$$

where the input is denoted h^0 and the output is given by x^L.

(a) **What is the Jacobian J of the input-output relation of this network?**

Hint: see eq. 2 of the paper.

Solution:
Using the chain rule gives:

$$J = \prod_{l=1}^{L} D^l W^l$$

where D^l is a matrix of pointwise derivatives of the nonlinearity ϕ at layer l:

$$(D^l)_{ij} = \frac{dx_j^l}{dh_i^l} = \phi'(h_i^l)\delta_{ij}. \tag{3}$$

(b) As the paper discusses, we are interested in the spectrum of singular values of J, but all of the tools we have developed so far deal with the eigenvalue spectrum.

In terms of the singular values of J, what are the eigenvalues of JJ^T?

The definition of dynamical isometry, the condition we're after, is that the magnitude of the singular values of J should concentrate around 1.

What is the dynamical isometry condition in terms of the eigenvalues of JJ^T?

Solution: The singular values of a matrix A are the square roots of the eigenvalues of AA^T, so the eigenvalues of JJ^T are the squared singular values of J.
Quick proof: By SVD,

$$A = U\Sigma V^\dagger$$

$$AA^T = (U\Sigma V^\dagger)(U\Sigma V^\dagger)^\dagger = U\Sigma^T\Sigma U^\dagger = U\Sigma^2 U^\dagger$$

where Σ^2 is composed of squared singular values and $V^\dagger$ is matrix V's conjugate transpose. Note that Σ^2 equals matrix D from a spectral decomposition of AA^T, which contains eigenvalues of AA^T. Thus the squared singular values of A equal the eigenvalues of AA^T. □
So, the dynamical isometry condition is that the spectrum of eigenvalues of JJ^T concentrates around unity.

(c) Now that we're focused on JJ^T instead of J, read the following section reproduced from the main paper, about the S-transform of JJ^T's spectral density:

$$S_{JJ^T} = \prod_{l=1}^{L} S_{W_l W_l^T} S_{D_l^2} = S_{WW^T}^L S_{D^2}^L,$$

where we have used the identical distribution of the weights to define $S_{WW^T} = S_{W_l W_l^T}$ for all l, and we have also used the fact the pre-activations are distributed independently of depth as $h_l \sim \mathcal{N}(0, q^*)$, which implies that $S_{D_l^2} = S_{D^2}$ for all l.Eqn. (12) provides a method to compute the spectrum $\rho_{JJ^T}(\lambda)$. Starting from $\rho_{W^TW}(\lambda)$ and ρ_{D^2}, we compute their respective S-transforms through the sequence of equations eqns. (7), (9), and (10), take the product in eqn. (12), and then reverse the sequence of steps to go from S_{JJ^T} to $\rho_{JJ^T}(\lambda)$ through the inverses of eqns. (10), (9), and (8). Thus we must calculate the S-transforms of WW^T and D^2, which we attack next for specific nonlinearities and weight ensembles in the following sections. In principle, this procedure can be carried out numerically for an arbitrary choice of nonlinearity, but we postpone this investigation to future work.

Prove the equation at the top of the box.

Hint: this is done in the first appendix of the paper. Note that you should assume free independence of the D's and W's.

The upshot of this problem is that we need to calculate the quantities S_{WW^T} and S_{D^2} for whatever nonlinearities and weight initialization schemes we're interested in.

Solution:

$$JJ^T = \left(\prod_{l=1} D^l W^l\right)\left(\prod_{l=1} D^l W^l\right)^T = \left(D_L W_L \dots D_1 W_1\right)\left(D_L W_L \dots D_1 W_1\right)^T$$

by expanding the product. So

$$S_{JJ^T} = S_{(D_L W_L \dots D_1 W_1)(D_L W_L \dots D_1 W_1)^T}$$

Since the S-transform is defined in terms of moments of the eigenvalue distribution, it is invariant to cyclic permutations (since the trace, which defines moments, is invariant to cyclic permutations). So, we can re-order matrices in the product, yielding:

$$S_{JJ^T} = S_{(W_L^T D_L^T D_L W_L)(D_{L-1} W_{L-1} \dots D_1 W_1)(D_{L-1} W_{L-1} \dots D_1 W_1)^T}$$

Then, assuming free independence, the S-transforms multiply:

$$S_{JJ^T} = S_{(W_L^T D_L^T D_L W_L)} S_{(D_{L-1} W_{L-1} \dots D_1 W_1)(D_{L-1} W_{L-1} \dots D_1 W_1)^T}.$$

Again using invariance to cyclic permutations:

$$S_{JJ^T} = S_{(D_L^T D_L W_L W_L^T)} S_{(D_{L-1} W_{L-1} \dots D_1 W_1)(D_{L-1} W_{L-1} \dots D_1 W_1)^T}.$$

And again assuming free independence:

$$S_{JJ^T} = S_{D_L^T D_L} S_{W_L W_L^T} S_{(D_{L-1} W_{L-1} \dots D_1 W_1)(D_{L-1} W_{L-1} \dots D_1 W_1)^T}.$$

Since D is diagonal,

$$S_{JJ^T} = S_{D_L^2} S_{W_L W_L^T} S_{(D_{L-1} W_{L-1} \dots D_1 W_1)(D_{L-1} W_{L-1} \dots D_1 W_1)^T}.$$

Continuing this procedure we get

$$S_{JJ^T} = \prod_{l=1}^{L} S_{D_l^2} S_{W_l^T W_l}$$

Since the weight matrices W^l for each layer are identically distributed, their S-transforms are equal, so we can drop the subscript and write:

$$S_{JJ^T} = S_{W^T W}^L \prod_{l=1}^{L} S_{D_l^2}$$

Finally, using the fact that D^l matrices are identically distributed gives the desired expression.

$$S_{JJ^T} = \left(S_{W^T W}\right)^L \left(S_{D^2}\right)^L$$

Problem 2: Warm up - Dynamical isometry conditions for a purely linear network

Relevant readings: Resurrecting the Sigmoid, Section 2.4

As a warm up, let's assume all nonlinearities are identity functions; that is, we have a purely linear network.

(a) **For our purely linear network, what is the Jacobian between the output, x^L, and the input, h^0, in terms of the weight matrices W^l and bias vectors b^l for each layer?**

Hint: compare equation (2) from the paper.

Solution:
From problem 1 we know that that $J = \prod_{l=1}^{L} D^l W^l$. Since we are using a linear network, $\phi(x) = x$, thus D is the identity matrix and $J = \prod_{l=1}^{L} W^l$.

(b) In the linear activation function case, we will consider two different ways of initializing the weight matrix W. Equivalently, we take W to be drawn from one of two random matrix ensembles:

- Random Gaussian weights in which each W_{ij}^l is drawn i.i.d. from a Gaussian with variance σ_w^2/N
- Orthogonal weight matrices, drawn from a uniform distribution over scaled orthogonal matrices obeying $(W^l)^T W^l = \sigma_W^2 I$.

In the orthogonal case, what is the eigenvalue spectrum of JJ^T in terms of σ_w. Is it possible to choose σ_w to ensure dynamical isometry, and if so, how?

Solution: Since J is the product of orthogonal matrices (up to the factor σ_W^2, it will itself be orthogonal (up to the factor σ_W^{2L}). Thus the singular values are all given by σ_W^L (since here $JJ^T = \sigma_W^{2L} I$).
So, we can achieve dynamical isometry by setting $\sigma_W = 1$.
In a way, this is cheating a bit. Dynamical isometry is the condition that all the singular values of the Jacobian square to unity, meaning that the Jacobian is a unitary (or orthogonal) matrix. When we take an orthogonal network with $\sigma_W = 1$, we're making the network itself just multiplication by an orthogonal matrix (and then adding some constant which is a combination of weight matrices and bias vectors). So of course it will have dynamical isometry.

(c) Now let's consider the random Gaussian case. Here the dynamical isometry condition is not as simple to calculate, but we can get a handle on it a few ways.

First, argue that requiring the mean squared singular value of J to be 1 (which is required for dynamical isometry) in this case means that $\sigma_w^2 = 1$.

Hint: recall the relation between χ and the mean-squared singular value of the layerwise Jacobian which you worked out in problem set 2.)

Unfortunately, it's tough to go further than that for the Gaussian case. In eqn. (13) of the main paper, which comes from reference 17, the spectral density is given exactly, without derivation. **Taking eqn. (13) as given, show either numerically or analytically that the maximum eigenvalue of JJ^T scales linearly with the number of layers, L, as L grows large.**

What does the above result imply for achieving dynamical isometry in linear networks with Gaussian initialization?

Solution:
As you worked out in problem set 2, the quantity χ, which was equal to the mean-squared singular value of the layerwise Jacobian, was given by (see also eqn. (5) of the paper)

$$\chi = \sigma_w^2 \int \mathcal{D}h \phi'(\sqrt{q^*}h)^2.$$

For a linear network the derivative is always unity so this equation just reduces to $\chi = \sigma_W^2$.
So for the mean-squared singular value to be 1, we need $\sigma_W = 1$.
To understand $O(L)$ growth of the maximum eigenvalue of JJ^T, see the solution to equation (13) from the paper (at the end of section 2.4).
Because the maximum eigenvalue scales linearly with L, it is impossible to achieve dynamical isometry in the large L limit.

Problem 3: S_{D^2} for ReLU and hard-tanh networks

Relevant readings: *Resurrecting the Sigmoid*, Sections 2.2 and 2.5

In this problem, we turn to networks with nonlinearities. We look at two nonlinearities here, the ReLU function and a piecewise approximation to the sigmoid known as the hard-tanh. These functions are defined as follows:

$$f_{\text{ReLU}}(x) = \begin{cases} 0 & x \leq 0 \\ x & x \geq 0 \end{cases} \tag{4}$$

$$f_{\text{HardTanh}}(x) = \begin{cases} -1 & x \leq -1 \\ x & -1 \leq x \leq 1 \\ 1 & x \geq 1 \end{cases} \tag{5}$$

We want the spectral density, $\rho_{JJ^T}(\lambda)$, of JJ^T, where J is the Jacobian. We will find this by first calculating its S-transform, S_{JJ^T}. As discussed in the introduction, this involves two separate steps: finding S_{D^2} and finding S_{WW^T}. Note that finding S_{D^2}'s closed form relies primarily on choice of nonlinearity, and finding S_{WW^T}'s closed form relies only on choice of weight initialization (and not on choice of nonlinearity).

In this problem, we focus on the nonlinearities (S_{D^2}); the next problems focus on the weight initializations (S_{WW^T}), and how to combine these to get the S transform of the Jacobian.

(a) The probability density function of the D matrix depends on the distributions of inputs to the nonlinearity. To calculate this, we will make a couple simplifying assumptions. The first assumption is that we initialize the network at a critical point (defined in problem set 2).

If we are interested in finding conditions for achieving dynamical isometry, why is it a good assumption that the network is initialized at criticality?

> **Solution:** The criticality condition, $\chi = 1$, implies that the *mean* squared singular value of J, or equivalently that the mean eigenvalue of JJ^T, is unity. Dynamical isometry means that the *entire* spectrum of squared singular values of J is concentrated around unity. So criticality is a prerequisite for dynamical isometry.

(b) The second assumption we make in calculating the distribution of inputs to the nonlinearity is that the we have settled to a stationary point of the length map (the variance map).

Reread section 2.2 of *Resurrecting the Sigmoid*, and argue why this is also a good assumption.

> **Solution:**
> As described in both *Exponential Expressivity in Deep Neural Networks Through Transient Chaos* and in section 2.2 of *Resurrecting the Sigmoid*, the empirical distribution of network pre-activations approximates a 0-mean, q^l-variance Gaussian distribution in the large-width limit. The length map describing the evolution of q^l has a fixed point, which the papers show empirically is rapidly converged to. Because of this rapid convergence, it is natural to assume that only a few initial layers are not characterized by this variance, and that we can neglet them in computing the spectrum of the network's Jacobian. Conveniently, assuming we are at a fixed point makes D^2 is independent of l, greatly simplifying our analysis.

(c) To find the critical points of both the ReLU and hard-tanh networks, recall from problem set 2 that criticality was defined by the condition $\chi = 1$, where χ is defined in eqn. (5) of the main paper. As in the paper, define $p(q^*)$ as the probability, given the variance q^*, that a given neuron in a layer is in its linear (i.e. not constant) regime.

Show that $\chi = \sigma_w^2 p(q^*)$.

Hint: plug the nonlinearity into the equation for χ and reduce

Solution:

$$\chi = \sigma_w^2 \int Dh\phi'((\sqrt{}(q^*)h)^2$$

Where Dh is the standard Gaussian measure. Note that when $\phi' = 0$ (the slope of the activation function is zero) then $\chi = 0$. Thus, since χ only takes on values in $0, 1$. Thus the Gaussian measure integral, which represents probability that $\phi' \neq 0$ reduces to $p(q^*)$, the probability that $\phi' = 0$, so $\chi = \sigma_w^2 p(q^*)$.

(d) **In terms of $p(q^*)$, what is the spectral density $\rho_{D^2}(z)$ (for both ReLU and hard-tanh networks) of the eigenvalues of D^2?** .

Solution:
Bernoulli with parameter equal to the probability of being in the linear regime. The Dirac delta expresses the fact that both ReLU and hard-tanh are piecewise linear with sections at value 0, so their probability of being in the linear regime is a step function – it allows us to express a discrete pdf (in this case with two values, 0 and 1).

(e) **Following equations 7-10 in the main paper, derive the Stieltjes transform $G_{D^2}(z)$, the moment-generating function $M_{D^2}(z)$, and the S-transform $S_{D^2}(z)$ in terms of $p(q^*)$.** Note: This should be the same for both ReLU and hard-tanh networks.

Solution: Recall that:

$$\rho_{D^2}(z) = (1 - p(q^*))\delta(z) + p(q^*)\delta(z-1)$$

Then recall the definition

$$G_{D^2}(z) = \int_{\mathcal{R}} \frac{\rho_x(t)dt}{z-t} = \frac{\rho_x(0)}{z} + \frac{\rho_x(1)}{z-1} = \frac{1-p(q^*)}{z} + \frac{p(q^*)}{z-1}$$

Then

$$\begin{aligned} M_{D^2}(z) &= zG_{D^2}(z) - 1 \\ &= z\left(\frac{1-p(q^*)}{z} + \frac{p(q^*)}{z-1}\right) - 1 \\ &= -p(q^*) + \frac{zp(q^*)}{z-1} \\ &= p(q^*)\left(\frac{z}{z-1} - 1\right) \\ &= \frac{p(q^*)}{z-1} \end{aligned}$$

Next use the definition

$$S_{D^2}(z) = \frac{1+z}{zM_{D^2}^{-1}(z)}.$$

The inverse $M_{D^2}^{-1}(z)$ is $\frac{p(q^*)}{z} + 1$. Thus:

$$S_{D^2}(z) = \frac{1+z}{z\left(\frac{p(q^*)}{z} + 1\right)} = \frac{z+1}{z+p(q^*)}$$

(f) Now that we've calculated the transforms we wanted in terms of $p(q^*)$, let us see what the critical point (which determines q^* and $p(q^*)$) looks like for our two nonlinearity options.

For ReLU networks, what is $p(q^*)$? Show that this implies that the only critical point for ReLU networks is $(\sigma_w, \sigma_b) = (\sqrt{2}, 0)$.

Solution:
For ReLUs, the nonlinearity is half in the positive linear regime and half at 0. Assuming 0-mean symmetric activation distributions, the probability of being in the linear regime is $p(q^*) = \frac{1}{2}$.
Using the above result that $\chi = \sigma_w^2 p(q^*)$ immediately tells us that $\sigma_w^2 = 2$.
Using equation (4) in the *Resurrecting the Sigmoid* paper,

$$q^* = \sigma_w^2 \int \mathcal{D}h \ \phi(\sqrt{q^*}h)^2 + \sigma_b^2, \tag{6}$$

and using the fact that ϕ is a ReLU, we can write

$$q^* = q^* \sigma_w^2 \int_{h>0} \mathcal{D}h \ h^2 + \sigma_b^2. \tag{7}$$

Since the integrand is an even function, it can be evaluated easily

$$q^* = \frac{1}{2} q^* \sigma_w^2 \int \mathcal{D}h \ h^2 + \sigma_b^2. \tag{8}$$

The integral now is the variance of h, which is unity by construction, so we simply get

$$q^* = \frac{1}{2} q^* \sigma_w^2 + \sigma_b^2. \tag{9}$$

Plugging in $\sigma_w^2 = 2$ gives $q^* = q^* + \sigma_b^2$, meaning $\sigma_b^2 = 0$.

(g) For hard-tanh networks, the behavior is a bit more complex, but we can calculate it numerically. As we saw in problem set 2, for the smooth tanh network there is a 1D curve in the (σ_w, σ_b) plane which satisfies criticality. The same is true for the hard tanh network, as we'll now see. We are interested in three quantities, all of which are functions of σ_w and σ_b: q^*, $p(q^*)$, and χ.

We've already seen (in part (c) above) that if we know σ_w and $p(q^*)$, we can easily determine χ. It turns out that there is also a simple relation between q^* and $p(q^*)$.

Show that for the hard tanh network, $p(q^*) = \text{erf}(1/\sqrt{2q^*})$.

Now all that's left is to determine q^* as a function of σ_w and σ_b, and then we can get both q^* and $p(q^*)$. Remember that in problem set 2, you derived the relation

$$q^* = \sigma_w^2 \int \ \mathcal{D}h \ \phi(\sqrt{q^*}h)^2 + \sigma_b^2. \tag{10}$$

Use this relation to get an implicit expression for q^* in terms of σ_w and σ_b.

Using the three relations, and any programming language or numerical package of your choice, plot (in the σ_w, σ_z plane) the three quantities of interest, and identify the critical line $\chi = 1$.

Solution:
For hard-tanh, $p(q^*)$ is the probability that a normally distribution set of activations takes on values in hard-tanh's linear regime (recall this is between -1 and 1). Thus we integrate $\int_{-1}^{1} z dz$ where z is a zero-mean Gaussian with variance q^*. The integral of the Gaussian is given by the error function. The error function (denoted erf and defined as the integral of the standard Gaussian) is commonly defined without the leading factor $\frac{2}{\pi}$, so $\int z dz = erf(\sqrt{(}1/2q^*)$ (the parameter $1/2q^*$ is arrived at by substituting $t = h/\sqrt{2q^*}$). Thus $p(q^*) = erf(\sqrt{(}1/2q^*)$.

Problem 4: Can Gaussian initialization achieve dynamical isometry?

Relevant readings: Resurrecting the Sigmoid, section 2.5

In this problem, we will consider weights with a Gaussian initialization, and use the results from the previous problems to investigate whether dynamical isometry can be achieved for such nets over our two main activation functions of interest (ReLU and hard-tanh).

(a) As we've seen in the decomposition from the previous problems, the S-transform of $\mathbf{J}\mathbf{J}^T$ depends on the S-transform of D^2, which was computed above, and that of WW^T, which is a *Wishart random matrix*, i.e. the product of two random Gaussian matrices.

Prove that $S_{WW^T}(z) = \frac{1}{\sigma_w^2 \cdot (z+1)}$, using the following connection between the moments of a Wishart matrix and the Catalan numbers:

$$m_k = \frac{\sigma_w^{2k}}{k+1}\binom{2k}{k}$$

where m_k is the k^{th} moment of WW^T.

Solution. Given the moments, we can easily form the moment-generating function

$$M_{WW^T}(z) = \sum_{k=1}^{\infty} \frac{m_k}{z^k} = \sum_{k=1}^{\infty} \left(\frac{\sigma_w^2}{z}\right)^k \frac{1}{k+1}\binom{2k}{k} = \sum_{k=1}^{\infty} \left(\frac{\sigma_w^2}{z}\right)^k C_k$$

where C_k is the k^{th} Catalan number. So, we can now exploit the defining recurrence relation for the Catalan numbers, that $C_k = \sum_{j=0}^{k-1} C_j C_{k-j-1}$ (if you think of the k^{th} Catalan number as the number of ways to balance $2k$ parentheses, this recurrence is pretty intuitive). To start off, this recurrence starts with the C_0, though our MGF does not, and this might make the calculation more difficult; let's temporarily work with

$$f(x) := \sum_{k=0}^{\infty} \left(\frac{\sigma_w^2}{z}\right)^k C_k = 1 + M_{WW^T}(z)$$

Next, the recurrence is in a sum of products of Catalan numbers; specifically, products whose indices have a constant sum. Seeing as $f(x)$ is basically an infinitely long polynomial, and polynomial multiplication also involves such product sums, a good first attempt to apply this recurrence is to square our function. Indeed, we have:

$$f(x)^2 = \sum_{k=0}^{\infty}\sum_{j=0}^{\infty} \left(\frac{\sigma_w^2}{z}\right)^{k+j} C_k C_j$$

which after collecting like terms, is

$$f(x)^2 = \sum_{k=0}^{\infty} \left(\frac{\sigma_w^2}{z}\right)^k \sum_{j=0}^{k-1} C_j C_{k-j} = \sum_{k=0}^{\infty} \left(\frac{\sigma_w^2}{z}\right)^k C_{k+1}$$

Thus,

$$\frac{\sigma_w^2}{z} f(x)^2 = \frac{\sigma_w^2}{z}\left(M_{WW^T}(z)+1\right)^2 = \sum_{k=1}^{\infty} \left(\frac{\sigma_w^2}{z}\right)^k C_k = M_{WW^T}(z)$$

Solving the quadratic equation yields

$$M_{WW^T}(z) = \frac{z}{2\sigma_w^2} - 1 - \frac{1}{2}\sqrt{1 - \frac{4\sigma_w^2}{z}}$$

Now that we've reduced the MGF to a quadratic polynomial, inverting it is easy enough, and we are left with

$$M^{-1}_{WW^T}(z) = \sigma_w^2 \frac{(z+1)^2}{z}$$

$$S_{WW^T}(z) = \left(\sigma_w^2 \cdot (z+1)\right)^{-1}$$

(b) We now have enough pieces to begin attacking the calculation of the Jacobian singular value distribution - recall that due to the decomposition

$$S_{JJ^T} = (S_{WW^T})^L \cdot (S_{D^2})^L$$

once we've calculated the S-transforms for D^2 and WW^T, we can easily obtain the S-transform of $\mathbf{JJ}^T$.

Using your solution to the previous part and the calculation of S_{D^2} from the earlier problems, show that

$$S_{JJ^T} = \sigma_w^{-2L} \cdot (z + p(q^*))^{-L}.$$

Solution. We calculated S_{D^2} in part (e) of problem 3, showing that

$$S_{D^2}(z) = \frac{z+1}{z+p(q^*)}$$

And from the previous part we know that

$$S_{WW^T}(z) = \frac{1}{\sigma_w^2(z+1)},$$

so combining these gives

$$S_{JJ^T} = (S_{WW^T})^L (S_{D^2})^L = \left(\sigma_w^{-2}(1+z)^{-1}\right)^L \left(\frac{1+z}{z+p(q^*)}\right)^L = \sigma_w^{-2L}(z+p(q^*))^{-L}$$

(c) From the S-transform, one route to getting information about the spectrum of JJ^T is to compute the spectral density $\rho_{JJ^T}(\lambda)$. While that calculation is too involved, we can get the answer to the question of achieving dynamical isometry by a slightly more indirect route.

Use the S-transform you calculated above to calculate $M^{-1}_{JJ^T}$ (the inverse of the moment-generating function for $\mathbf{JJ}^T$).

Hint: To compute the inverse MGF, recall the definition of the S-transform given in the paper (section 2.3, eqn. 10).

Solution. The S-transform is defined so that $S_{JJ^T} = \frac{1+z}{zM^{-1}_{JJ^T}(z)}$, so

$$M^{-1}_{JJ^T}(z) = \frac{1+z}{zS_{JJ^T}(z)} = \frac{1+z}{z}\left(z+p(q^*)\right)^L \sigma_w^{2L}$$

(d) We can now compute the variance of the JJ^T eigenvalue distribution, $\sigma^2_{JJ^T}$. You should have calculated above that

$$M^{-1}_{JJ^T}(z) = \frac{1+z}{z} \cdot (z+p(q^*))^L \cdot \sigma_w^{2L}$$

Using the definition that

$$M_{JJ^T}(z) = \sum_{k=1}^{\infty} \frac{m_k}{z^k}$$

and the expression for the functional inverse of M_{JJ^T} to compute that the first two moments are

$$m_1 = \sigma_w^{2L} p(q^*)^L$$

$$m_2 = m_1^2 \cdot \frac{L+p(q^*)}{p(q^*)}$$

Hint: Use the Lagrange inversion theorem (eqn. 18 in the paper) to obtain a power series for the inverse MGF and equate corresponding coefficients with our calculated expressions.

Solution
Note that we have a formula for $M^{-1}(z)$ (suppressing the JJ^T subscript for clarity, but the moments are defined in terms of $M(z)$. In the paper, the Lagrange inversion theorem is used to express the constant and $1/z$ coefficients of $M^{-1}(z)$ in terms of the m_1 and m_2. Here is a slightly hand-wavy proof of that result (the rigorous proof turns out to be quite difficult):

Proof. Assume that the $M^{-1}(z)$ can be written as a Taylor series with an additional $1/z$ term (This assumption is one of the weaknesses of this proof). So

$$M^{-1}(z) = \frac{a}{z} + b + cz + dz^2 + \cdots \tag{11}$$

Since we know that

$$M(z) = \frac{m_1}{z} + \frac{m_2}{z^2} + \cdots, \tag{12}$$

we can write

$$z = \frac{m_1}{M^{-1}(z)} + \frac{m_2}{M^{-1}(z)^2} + \cdots \tag{13}$$

Plugging in our ansatz above gives

$$z = \frac{m_1}{\left(\frac{a}{z} + b + cz + dz^2 + \cdots\right)} + \frac{m_2}{\left(\frac{a}{z} + b + cz + dz^2 + \cdots\right)^2} + \cdots \tag{14}$$

We'll expand the RHS of the above equations assuming z to be small, and then equate coefficients of the RHS and LHS. Specifically, we will expand the RHS to second order in z.

$$z = \frac{m_1 z}{a}\left(1 + (b/a) + (c/a)z + (d/a)z^2 + \cdots\right)^{-1} +$$

$$\frac{m_2 z^2}{a^2}\left(1 + (b/a) + (c/a)z + (d/a)z^2 + \cdots\right)^{-2} + \cdots = \frac{m_1}{a}z - \frac{m_1 b}{a^2}z^2 + \frac{m_2}{a^2}z^2 + O(z^3)$$

Since the coefficient of z above has to be unity, and the coefficient of z^2 has to be zero, this implies that $a = m_1$ and $b. = m_2/m_1$. This implies that our sought-after expression for $M^{-1}(z)$ is

$$M^{-1}(z) = \frac{m_1}{z} + \frac{m_2}{m_1} + \cdots \tag{15}$$

□

With this expression in hand, we can directly extract the constant and $1/z$ coefficients of the function $M^{-1}(z)$:
Given the result of our earlier calculation that

$$M^{-1}_{JJ^T}(z) = \left(1 + \frac{1}{z}\right) \cdot (z + p(q^*))^L \cdot \sigma_w^{2L},$$

we see that the only place to get a $1/z$ term here is from the constant term when the $(z + p(q^*))^L$ is expanded. This constant term will simply be $p(q^*)^L$, so the $1/z$ term here, which is m_1, is

$$m_1 = \sigma_w^{2L} p(q^*)^L \tag{16}$$

The constant term comes from two places. One, the $p(q^*)^L$ multiplies the 1 in the first term, and the term $Lzp(q^*)^{L-1}$, coming from the binomial expansion, multiplies the $1/z$ in the first term. So this means that the constant coefficient, m_2/m_1, is given by

$$\frac{m_2}{m_1} = \sigma_w^{2L} p(q^*)^L + \sigma_w^{2L} L p(q^*)^{L-1}. \tag{17}$$

Recognizing the first term in the RHS sum as m_1, we can factor to get

$$\frac{m_2}{m_1} = m_1\left(1 + \frac{L}{p(q^*)}\right), \tag{18}$$

or,

$$m_2 = m_1^2\left(1 + \frac{L}{p(q^*)}\right), \tag{19}$$

as desired.

(e) **From the first two moments calculated above, calculate the variance $\sigma^2_{JJ^T}$.**

Show that this variance scales linearly with depth, and **argue why this scaling implies that dynamical isometry can never be achieved in Gaussian networks.**

Solution

For any probability distribution, the variance is the second moment minus the first moment squared, so

$$\sigma^2_{JJ^T} = m_2 - m_1^2. \tag{20}$$

Plugging in the expressions we derived above gives

$$\begin{aligned} \sigma^2_{JJ^T} &= m_1^2 \left(1 + \frac{L}{p(q^*)}\right) - m_1^2 && (21) \\ &= m_1^2 \frac{L}{p(q^*)}. && (22) \end{aligned}$$

The above expression holds in general, but note that the condition for criticality, which as discussed above is a prerequisite for dynamical isometry, is simply that the mean eigenvalue of JJ^T, or equivalently the mean-squared singular value of J, is unity. Thus criticality implies that $m_1 = 1$, and so the variance of the eigenvalue distribution at criticality is given by

$$\sigma^2_{JJ^T,\text{crit}} = \frac{L}{p(q^*)}. \tag{23}$$

For ReLU networks, where $p(q^*)$ is always $1/2$, this reduces to $2L$, but no matter what the nonlinearity, since $p(q^*)$ is always less than unity, the variance of the eigenvalue distribution is always greater than L. So as the networks as made deeper and deeper, there is no way to achieve the dynamical isometry condition, which required that all the eigenvalues of JJ^T concentrate around unity. Clearly, this is incompatible with a variance which grows linearly with the depth of the network.

So, the unfortunate conclusion of this problem is that no matter what nonlinearity is chosen, **we cannot achieve dynamical isometry in networks where we utilize Gaussian initialization.**

Problem 5: Can orthogonal initialization achieve dynamical isometry?

Relevant readings: Resurrecting the Sigmoid, section 2.5

Having seen that Gaussian initialization is not capable of achieving dynamical isometry—no matter what the nonlinearity—we now ask whether it is possible to achieve dynamical isometry with an orthogonal initialization.

(a) **Compute the S-transform of WW^T for the orthogonal case, using the fact that for the orthogonal initialization, $WW^T = \sigma_w^2 I$.**

Solution. By definition, $WW^T = \sigma_w^2 I$. Scaling a matrix scales its moments, and hence inversely scales its S-transform, so we can first compute the S-transform for the identity and then scale appropriately. Since the identity matrix has only one eigenvalue - 1 - with maximal multiplicity, all the moments are 1 which means

$$M_I(z) = \sum_{k=1}^{\infty} z^{-k} = \frac{1}{z-1}$$

and so

$$M_I^{-1}(z) = \frac{1+z}{z}$$

We can now easily compute the S-transform of the identity as

$$S_I = \frac{1+z}{z} \cdot \frac{z}{1+z} = 1.$$

Doing the same thing for the S-transform of the identity scaled by σ_w^2 gives

$$M_{\sigma_w^2 I}(z) = \frac{1}{\frac{z}{\sigma_w^2} - 1},$$

meaning that the inverse MGF is now

$$M_{\sigma_w^2 I}^{-1}(z) = \sigma_w^2 \left(\frac{z+1}{z}\right),$$

which means that the S transform scales inversely by σ_w^2: $S_{WW^T} = S_{\sigma_w^2 I} = \sigma_w^{-2}$.

(b) **Combine your results above with previous calculations to arrive at the following expressions for the S-transforms and inverse moment-generating-functions for this orthogonal case:**

$$S_{JJ^T}(z) = \sigma_w^{-2L} \left(\frac{z+1}{z+p(q^*)}\right)^L, \tag{24}$$

$$M_{JJ^T}^{-1}(z) = \frac{z+1}{z} \left(\frac{z+1}{z+p(q^*)}\right)^{-L} \sigma_w^{2L} \tag{25}$$

Solution. We know that

$$S_{JJ^T} = S_{WW^T}^L S_{D^2}^L$$

and

$$S_{D^2} = \frac{z+1}{z+p(q^*)}$$

So it follows that for orthogonal weight initialization,

$$S_{JJ^T} = \sigma_w^{-2L} \cdot \left(\frac{z+1}{z+p(q^*)}\right)^L$$

From definition of the S-transform, we then have

$$M_{JJ^T}^{-1}(z) = \frac{z+1}{z} \cdot (S_{JJ^T})^{-1} = \frac{z+1}{z} \left(\frac{z+1}{z+p(q^*)}\right)^{-L} \sigma_w^{2L}$$

(c) Now we again look for the variance $\sigma_{JJ^T}^2$, using the Lagrange inversion theorem which you used in the previous problem.

Show that at criticality,

$$\sigma_{JJ^T}^2 = \frac{1 - p(q^*)}{p(q^*)} L \tag{26}$$

for large L.

Solution. We found in a previous problem that, upon applying the Lagrange inversion theorem, the inverse of the moment-generating function is a power series starting with

$$M_{JJ^T}^{-1}(z) = \frac{m_1}{z} + \frac{m_2}{m_1} + \cdots$$

where m_1, m_2 are the first two moments. If we expand $M_{JJ^T}^{-1}(z)$ as a Laurent series and equate the first two coefficients, we'll be able to deduce the first two moments. Equivalently, expanding the first two terms of the Taylor series of

$$zM_{JJ^T}^{-1}(z) = m_1 + \frac{m_2}{m_1} z + \cdots = (z+1)\left(\frac{z+1}{z + p(q^*)}\right)^{-L} \sigma_w^{2L}$$

we find that

$$(z+1)\left(\frac{z+1}{z + p(q^*)}\right)^{-L} \sigma_w^{2L} = p(q^*)^L \sigma_w^{2L} + z \cdot p(q^*)^{L-1} \sigma_w^{2L} \cdot (L + p(q^*) - Lp(q^*)) + \cdots$$

Equating coefficients, we have

$$m_1 = (p(q^*)\sigma_w^2)^L$$

$$m_2 = \frac{m_1^2 \cdot (L + p(q^*) - Lp(q^*))}{p(q^*)} = m_1^2 \cdot \left(1 + \frac{L - Lp(q^*)}{p(q^*)}\right)$$

At criticality, $\chi = \sigma_w^2 p(q^*) = 1$, and so the variance is

$$\sigma_{JJ^T}^2 = m_2 - m_1^2 = m_1^2 \cdot \left(1 + \frac{L - Lp(q^*)}{p(q^*)}\right) - m_1^2 = \frac{L - Lp(q^*)}{p(q^*)} = \frac{1 - p(q^*)}{p(q^*)} L$$

as desired.

(d) **What does the above expression become in the case of a ReLU network? What does this imply for achieving dynamical isometry in such a network?**

Since for a ReLU network $p(q^*) = 1 - p(q^*) = 1/2$, the expression becomes $\sigma_{JJ^T} = L$. Thus, as for the case of Gaussian initialization, we find that the variance of the eigenvalue distribution of JJ^T grows without bound as we increase the number of layers, regardless of how we initialize the network, **so we cannot achieve dynamical isometry in this sort of network.**
Combined with the results of the previous problem, this shows that **we cannot achieve dynamical isometry in ReLU networks with either i.i.d. Gaussian or orthogonal initialization.**

(e) In a hard-tanh network, you showed that at criticality, a range of values of q^*, and therefore $p(q^*)$, were achievable, depending on the exact values of σ_w and σ_b.

Argue that no matter what the number of layers L is, we can always choose an initialization to make $\sigma_{JJ^T}^2$ as small as we want, and hence, can achieve dynamical isometry. How exactly do we do this?

The magnitude of $\sigma^2_{JJ^T}$ depends entirely on the factor

$$\frac{1-p(q^*)}{p(q^*)}, p(q^*) \in [0,1]$$

Notice that as $p(q^*)$ approaches 1, this factor approaches 0, and so if we can figure out how to control $p(q^*)$ we can make the variance arbitrarily small. In hard-tanh networks, we showed in a previous problem that $p(q^*) = \text{erf}\left((2q^*)^{-\frac{1}{2}}\right)$, which is continuous in q^*. To push the error function towards 1, we have to increase its argument without bound, which means decreasing q^*. Since q^* depends continuously on σ_w and σ_b, even if we restrict ourselves to the level curve where $\chi(\sigma_w, \sigma_b) = 1$, by pushing σ_w towards 1 and σ_b towards 0, we can control q^* and thus, by extension, σ_{JJ^T}. See figure 1 in the paper for an illustration of the dependence of q^* on σ_w and σ_b.