Problem Set 1: Signal propagation in the large-width limit

Depth First Learning Week 2

Problem 1: NN Signal propagation framework

In this problem, we will work with the basic framework for analyzing signal propagation in a feedforward neural networks as the width of the network's hidden layers grows towards infinity.

To set up the notation, let $x^0 \in \mathbb{R}^N$ be the (vector-valued) input to the network. Each layer, l, of the network, is a function from $\mathbb{R}^N \to \mathbb{R}^N$ defined via the two equations:

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

Here x^{l-1} is the input to the layer, x^l is the output, W^l is an $N \times N$ matrix containing the layer's weights, and b^l is an N-dimensional vector containing the layer's biases. The function ϕ is the nonlinearity (e.g., sigmoid, ReLU, etc.) used by the network. We sometimes call the inputs to the nonlinearity (h^l) the *pre-activations*, and we call the outputs of the nonlinearity (which are also the outputs of the layer) the *activations*.

Note here that we are considering a specific neural network architecture in which the number of hidden units does not vary from layer to layer.

(a) To initialize the network, we usually draw the weights and biases randomly from some distribution. Let's ignore the bias term for now (i.e. set $b^l = 0$ for all layers l). Since the pre-activations h^l are functions of the random variables, they themselves are (vector-valued, N-dimensional) random variables. We want to understand what we can say about the distribution of the h^l's as we move through layers of the network, i.e. as a function of l.

Suppose we initialize each weight matrix W^l from a zero-mean Gaussian with variance σ^2, i.e. $W^l \sim \mathcal{N}(0, \sigma^2)$. For simplicity, assume the nonlinearity is just the identity function; this way, we can study the special case of purely linear networks, similar to linear regression. We'll relax this assumption in subsequent problems.

What are the mean and variance of the distribution of a single-component, h_i^l, in terms of the mean and the variance of the the previous layer's preactivations, h_i^{l-1}?

Solution:
First calculate the mean, denoted $\langle h_i^l \rangle$:

$$\begin{aligned}
h_i^l = \sum_j W_{ij}^l x_j^{l-1} \rightarrow \langle h_i^l \rangle &= \sum_j \langle W_{ij}^l x_j^{l-1} \rangle && \text{(linearity of expectation)} \\
\rightarrow \langle h_i^l \rangle &= \sum_j \langle W_{ij}^l \rangle \langle x_j^{l-1} \rangle && (W \text{ independent of } x) \\
\rightarrow \langle h_i^l \rangle &= 0 && (W \text{ is zero-mean})
\end{aligned}$$

Now that the mean is known to be zero, we simply have to calculate the second moment $\langle (h_i^l)^2 \rangle$ to get the variance:

$$\begin{aligned}
\langle (h_i^l)^2 \rangle &= \left\langle \left(\sum_j W_{ij}^l x_j^{l-1} \right)^2 \right\rangle \\
&= \left\langle \sum_{jj'} W_{ij}^l W_{ij'}^l x_j^{l-1} x_{j'}^{l-1} \right\rangle \\
&= \sum_{jj'} \langle W_{ij}^l W_{ij'}^l \rangle \langle x_j^{l-1} x_{j'}^{l-1} \rangle && (W \text{ and } x \text{ independent}) \\
&= \sum_j \langle (W_{ij}^l)^2 \rangle \langle (x_j^{l-1})^2 \rangle && \text{(entries of } W \text{ independent)} \\
&= \sigma^2 \sum_j \langle (x_j^{l-1})^2 \rangle && \text{(weights have unit variance)} \\
&= \sigma^2 \sum_j \langle (h_j^{l-1})^2 \rangle && \text{(assumption that nonlinearity is just the identity)} \\
&= N\sigma^2 \text{Var}(h_i^{l-1}) && \text{(symmetry)}
\end{aligned}$$

So we find that the variance of any given neuron's pre-activation grows (or shrinks) by a factor of $N\sigma^2$ at each layer.

(b) You should find in the previous part that the mean of the distribution of h_i^l is always zero, but that each layer of the network multiplies the variance of the distribution by a factor of $N\sigma^2$. Typically, the variance of the distribution of pre-activations is a proxy for how much of the nonlinearity we're making use of (e.g. for the sigmoid, if the variance is very small, we're basically in the linear region of the nonlinearity). To be able to vary the number of layers in a network and not have the behavior change too much, we'd like to have an initialization strategy which keeps the variance of the h^l distribution the same as we change the number of layers. Clearly, our current strategy ($W^l \sim \mathcal{N}(0, \sigma^2)$) doesn't accomplish this.

Suggest a simple modification of the initialization strategy which would achieve this goal.

Solution Instead of initializing with a fixed variance σ^2 which doesn't depend on the number of hidden units in the layer, we can pick some variance σ_w^2 and scale this variance by the number of hidden units in the layer, i.e. we take the initialization strategy $W^l \sim \mathcal{N}(0, \sigma_w^2/N)$.

(c) Now let's add back in the bias term. Imagine that we initialize according to a simple zero-mean Gaussian, where the variance (as for the weights in part (a)) was independent of the number of hidden units N, that is $b^l \sim \mathcal{N}(0, \sigma_b^2)$.

Does this initialization suffer from the same problem as the weight initialization described in part (b)? If so, how can we fix it?

Solution No, in this case the problem is not there, because each neuron's pre-activation only depends on a single bias term, i.e.

$$\langle (h_i^l)^2 \rangle = \langle (W_{ij}^l x_j^{l-1})^2 \rangle + \langle (b_i^l)^2 \rangle \tag{3}$$

And this does not change in the same way as the number of layers grows.

Problem 2: $N \to \infty$ and the mean-field approximation

In this problem, we use the knowledge we gained in problem 1 to properly choose to initialize the weights and biases according to $W^l \sim \mathcal{N}(0, \sigma_w^2/N)$ and $b^l \sim \mathcal{N}(0, \sigma_b^2)$. We'll investigate some techniques that will be useful in understanding precisely how the network's random initialization influences what the net does to its inputs; specifically, we'll be able to take a look at how the *depth* of the network together with the initialization governs the propagation of an input point as it flows forward through the network's layers.

(a) **A natural property of input points to study as the input flows through the net layer by layer is its length. Intuitively, this is closely related to how the net transforms the input space, and to how the depth of the net relates to that transformation. Compute the length q^l of the activation vector outputted by layer l. When considering non-rectangular nets, where layer l has length N_l, we want to distinguish this activation norm from the width of individual layers, so what's a more appropriate quantity we can track to understand how the lengths of activation vectors change in the net?**

Solution The length is simply the Euclidean magnitude, i.e. $\sum_{i=1}^{N} (h_i^l)^2$. We can stabilize this quantity, especially when N differs across layers, by normalizing:

$$q^l = \frac{1}{N_l} \sum_{i=1}^{N_l} (h_i^l)^2$$

(b) **What probabilistic quantity of the neuronal activations does q^l approximate (with the approximation improving for larger N)?**

Hint: recall that all neuronal activations h_i^l are zero-mean, and consider the definition of q^l from part (a) in terms of the empirical distribution of h_i^l.

Solution
q^l is the second moment of the empirical distribution of layer l activations, and hence approximates the variance. Indeed, as $N \to \infty$, the empirical average can be written $q^l = \mathbb{E}\left((h_i^l)^2\right) = \mathrm{Var}(h_i^l)$.

(c) **Calculate the variance of an individual neuron's pre-activations, that is, the variance of h_i^l. Your answer should be a recurrence relation, expressing this variance in terms of h^{l-1} (and the parameters σ_w and σ_b).**

(Note: You basically did this in problem 1; the differences here are just that the weights are initialized slightly differently and that the bias term exists, and now the noninearity is not just he identity)

Solution
Because the means of both the weight and bias distributions are zero, to calculate the variance we just need to calculate the second moment. We can use the fact that the weights and biases are initialized independently, so that the variance of h_i^l is the sum of a bias term and a variance term:

$$\begin{aligned}
\langle (h_i^l)^2 \rangle &= \left\langle \left(\sum_j W_{ij}^l x_j^{l-1} \right)^2 \right\rangle \\
&= \left\langle \sum_{jj'} W_{ij}^l W_{ij'}^l x_j^{l-1} x_{j'}^{l-1} \right\rangle + \langle (b_i^l)^2 \rangle \\
&= \left\langle \sum_{jj'} W_{ij}^l W_{ij'}^l x_j^{l-1} x_{j'}^{l-1} \right\rangle + \sigma_b^2 \\
&= \frac{\sigma_w^2}{N} \sum_j \langle (x_j^{l-1})^2 \rangle + \sigma_b^2 \\
&= \sigma_w^2 \langle (x^{l-1})^2 \rangle + \sigma_b^2 \\
&= \sigma_w^2 \langle \phi(h^{l-1})^2 \rangle + \sigma_b^2
\end{aligned}$$

(d) Now consider the limit that the number of hidden neurons, N, approaches infinity.

Use the central limit theorem to argue that in this limit, the pre-activations will be zero-mean Gaussian distributed. Be explicit about the conditions under which this result holds.

Solution
The basic idea here is to use the central limit theorem since the pre-activation is a sum of a large number of random variables, i.e.:

$$h_i^l = \sum_j^N W_{ij}^l x_j^{l-1} + b_i^l. \tag{4}$$

There are N terms in the sum, so as N goes to infinity, we should have a sum of a large number of random variables which should be well-approximated by a Gaussian.
However, there are a few things we need to be careful of:

i CLT can show that the sum $\sum_j^N W_{ij}^l x_j^{l-1}$ is Gaussian-distributed, but there is still the bias term b_i^l. So we do have to assume that the bias term is Gaussian-distributed as well.

ii In order to use CLT, we need each of the variables being added to have finite variance. These individual variables are $W_{ij}^l x_j^{l-1}$. By construction the weights have finite variance; what about the previous layer's activations, x_j^{l-1}? Unless the activation function ϕ is pathological, if we *assume that the previous-layer pre-activations have finite variance*, there should not be a problem here. In fact, if we just assume that the input distribution, i.e. x^0, has finite variance, all the layers' activations do too. Certainly the commonly used activation functions sigmoid, ReLU, etc. cannot turn a finite-variance sample of pre-activations into an infinite-variance sample of activations.

iii In order to use the CLT, we also need each of the variables being added to have identical distributions. This is true by symmetry.

iv The final condition for use of CLT is that the variables being added are all independent. Taking another look at the definition of q^l,

$$q^l = \frac{1}{N}\sum_{i=1}^N (h_i^l)^2$$

we want to show that each h_i^l is independent (from which the independence of their squares follows). Each h_i^l is in turn defined

$$h_i^l = \sum_{j=1}^N W_{ij}^l \phi(h_j^{l-1}) + b_i^l$$

By assumption, W_{ij}^l and b_i^l are independent from each other and, over all i, j, from any quantities in previous layers, including $\phi(h_j^{l-1})$. But are the W_{ij}^l independent of the h_j^{l-1}? To justify this, observe that we can view the sum above as a linear combination of the random variables W_{ij}^l; even though, technically, the linear combination is also over random variables $\phi(h_j^{l-1})$, the key is that over $1 \leq i \leq N$, all the h_j^{l-1}'s are the same. In other words, each neuronal activation in layer l depends on the same exact realization of the random variables that are the activations of the previous layer. So, h_i^l is essentially a linear combination of the (independent) W_{ij}^l with deterministic weights, at least with respect to i. So, we can justify the use of the CLT in analyzing $\lim_{N\to\infty} q^l$.

(e) With this zero-mean Gaussian approximation of q^l, we have a single parameter characterizing this aspect of signal propagation in the net: the variance, q^l, of individual neuronal activations (a proxy for squared activation vector lengths). Let's now look at how this variance changes from layer to layer, by deriving the relationship between q^l and q^{l-1}.

In part (c), your answer should have included a term $\langle (x^{l-1})^2 \rangle$. In terms of the activation function ϕ and the variance q^{l-1}, write this expectation value as an integral over

the standard Gaussian measure.

Solution
Since $x_i^{l-1} = \phi(h_i^{l-1})$, we can write the variance $\langle (x^{l-1})^2 \rangle$ as

$$\begin{aligned}\langle (x^{l-1})^2 \rangle &= \langle \phi(h^{l-1})^2 \rangle \\ &= \int_{\mathbb{R}} dx\ \phi(x)^2\ p_{h^{l-1}}(x),\end{aligned}$$

where $p_{h^{l-1}}(x)$ is the pdf of the pre-activations h^{l-1}. By assumption this is a zero-mean Gaussian of variance q^{l-1}, i.e.

$$p_{h^{l-1}}(x) = \frac{1}{\sqrt{2\pi q^{l-1}}} e^{-\frac{x^2}{2q^{l-1}}}.$$

This can be written in terms of the standard Gaussian distribution $\rho(x)$ via the change of variables

$$p_{h^{l-1}}(x) = \frac{1}{\sqrt{q^{l-1}}} \rho(x/\sqrt{q^{l-1}}),$$

meaning that the variance $\langle (x^{l-1})^2 \rangle$ becomes

$$\langle (x^{l-1})^2 \rangle = \frac{1}{\sqrt{q^{l-1}}} \int_{\mathbb{R}} dx\ \phi(x)^2\ \rho(x/\sqrt{q^{l-1}}).$$

Let $y = x/\sqrt{q^{l-1}}$, then

$$\langle (x^{l-1})^2 \rangle = \int_{\mathbb{R}} dy\ \phi(y\sqrt{q^{l-1}})^2\ \rho(y).$$

Use this result to write a recursion relation for q^l in terms of q^{l-1}, σ_w, and σ_b.

Solution
We just plug in, to get

$$q^l = \sigma_w^2 \int_{\mathbb{R}} dy\ \phi(y\sqrt{q^{l-1}})^2\ \rho(y) + \sigma_b^2$$

Problem 3: Fixed points and stability

In the previous problem, we found a recurrence relation relating the length of a vector at layer l of a network to the length of the vector at the previous layer, $l-1$ of the network. In this problem, we are interested in studying the properties of this recurrence relation. In the *Resurrecting the sigmoid* paper, the results of this problem are used to understand at which bias point to evaluate the Jacobian of the input-output map of the network. For more information on this topic, see either of the two papers which are suggested reading for this week:

- *Exponential expressivity in deep neural networks through transient chaos*
- *Deep information propagation*

Note that in this problem, we are just taking the recurrence relation as a given, i.e. we do not need to worry about random variables or probabilities; all of that went into determining the recurrence relation. Instead, we'll use tools from the theory of dynamical systems to investigate the properties - in particular, the asymptotics - of this recurrence relation.

(a) A simple example of a dynamical system is a recurrence defined by some initial value x_0 and a relation $x_n = f(x_{n-1})$ for all $n > 0$. This system defines the resulting sequence x_n. Sometimes, these systems have *fixed points*, which are values x^* such that $f(x^*) = x^*$.

If the value of the system, x_m, at some time-step m, happens to be a fixed point x^*, what is the subsequent evolution of the system?

Solution
Since $f(x^*) = x^*$, for all times greater than m, the system simply stays at x^*.

(b) **For the recurrence relation you derived in the previous problem, what is the equation which a fixed-point of the variance, q^*, must satisfy?**

Under some conditions (i.e. for some values of σ_w and σ_b), the value $q^* = 0$ is a fixed point of the system. What are these conditions?

Solution
A fixed point has to satisfy

$$q^* = \sigma_w^2 \int_{\mathbb{R}} \phi\left(\rho\sqrt{q^*}\right)^2 \, \mathrm{d}\rho + \sigma_b^2 \tag{5}$$

where $\mathrm{d}\rho$ is the standard Gaussian measure. If $\sigma_b = 0$, i.e. there is no bias term, and the nonlinearity has a zero y-intercept, then there is a trivial fixed point of $q^* = 0$.

(c) Now let us be concrete, and look at the recurrence relation in the special case of a nonlinearity $\phi(h)$ which is both monotonically increasing and satisfies $\phi(0) = 0$. Note that both of the nonlinearities considered in the paper we are studying, the tanh and ReLU nonlinearities, satisfy this property.

Show that those two properties (monotonicity and $\phi(0) = 0$) imply that the length map $q^l(q^{l-1})$ is monotonically increasing.

Solution
To prove that the function is monotonically increasing with its argument q, we take the derivative:

$$f(q) = \sigma_w^2 \int_{\mathbb{R}} \phi(\rho\sqrt{q})^2 \, d\rho + \sigma_b^2$$
$$f'(q) = \frac{\sigma_w^2}{\sqrt{q}} \int_{\mathbb{R}} \phi(\rho\sqrt{q})\phi'(\rho\sqrt{q})\rho d\rho$$

The derivative is positive since by assumption ϕ' is positive everywhere, and $\phi\rho$ is also positive everywhere. So the function is monotonically increasing.
Optional: Show that these two properties imply that the length map $q^l(q^{l-1})$ is a concave function.
(Note: We have not managed to prove this ourselves (and not for lack of trying!), so feel free to skip)

Solution

What is the maximum number of times any concave function can intersect the line $y = x$? What does this imply about the number of fixed points the length map $q^l(q^{l-1})$ can have?

Solution
Note that since a fixed point is defined as a point, x^*, such that $f(x^*) = x^*$, graphically the fixed point can be found from the intersection of the length map $q^l(q^{l-1})$ with the line $y = x$.
If you think about the definition of a concave function (specifically, the version of the definition which states that beween any two points $x = a$ and $x = b$, the graph of the function must lie above the line defined by $f(a)$ and $f(b)$), you will realize that a concave function cannot intersect any line more than twice. Thus, concavity implies that the function can have at most two fixed points.

(d) Let's be concrete now and consider the nonlinearity to be a ReLU.

Compute (analytically) the length map $q^l = f(q^{l-1})$, which will also depend on σ_w and σ_b. For what values of σ_w and σ_b does the system have fixed point(s)? How does the value of the fixed point depend on σ_w and σ_b?

Solution
Starting from

$$f(q) = \sigma_w^2 \int_{\mathbb{R}} \phi(\rho\sqrt{q})^2 \, d\rho + \sigma_b^2, \tag{6}$$

and explicitly inserting the nonlinearity ϕ gives

$$f(q) = \sigma_w^2 \int_0^\infty \rho^2 q \, d\rho + \sigma_b^2. \tag{7}$$

Note that since the ReLU nonlinearity is zero when the argument is zero and just the identity function when the argument is greater than zero, we can take its effect into account simply by changing the above limits of integration so that we only integrate over the region in which the argument is positive. Now we can pull q out of the integral,

$$f(q) = q\sigma_w^2 \int_0^\infty \rho^2 \, d\rho + \sigma_b^2, \tag{8}$$

and to evaluate the integral, note that by symmetry of the Gaussian distribution, it's half of what it would be if we had the limits from $-\infty$ to ∞, in which case it would just be the variance of a standard Gaussian, and so

$$f(q) = q\frac{\sigma_w^2}{2} + \sigma_b^2. \tag{9}$$

The important things to note here are that because $f(q)$ is a simple linear function, there is at most a single fixed point of the system. If σ_b^2 is zero, that fixed point is at $q = 0$. If $\sigma_b^2 > 0$, then there is a fixed point only if $\sigma_w < \sqrt{2}$. Otherwise, the system does not have any fixed point. This is a qualitative difference from the tanh case, in which there is always a fixed point.
A slightly strange case is when $\sigma_w = \sqrt{2}$ exactly, and $\sigma_b = 0$. In this case, the recurrence relation gives $q^l(q^{l-1}) = q^{l-1}$, meaning that every point is a fixed point.

(e) Now let's consider the sigmoid nonlinearity $\phi(h) = \tanh(h)$. In this case the length map cannot be computed analytically, but it can be done numerically.

Numerically plot the length map, $q^l = f(q^{l-1})$, for a few values of σ_w and σ_b in the following regimes: (i) $\sigma_b = 0$ and $\sigma_w < 1$, (ii) $\sigma_b = 0$ and $\sigma_w > 1$, and (iii) $\sigma_b > 0$. Describe qualitatively the fixed points of the map in each regime.

Solution
The following Python code should work:

```
import numpy as np
import scipy.integrate as integrate

def integrand(x):
    gaussian = np.sqrt(2 * np.pi), -np.inf, np.inf) * np.exp(-0.5 * x**2)
    return np.tanh(x * np.sqrt(q))**2 * gaussian

def fint(q):
    result = integrate.quad(integrand)
    return result[0]

def lengthmap(q, sigma_w, sigma_b):
    return sigma_w**2 * fint(q) + sigma_b**2
```

The behavior that should be seen is the following, as described in the transient chaos paper (ignoring the parts about stability because we haven't covered that yet. See next part of the problem):
For $\sigma_b = 0$ and $\sigma_w < 1$, the only intersection is at $q^* = 0$. In this bias-free, small weight regime, the network shrinks all inputs to the origin. For $\sigma_w > 1$ and $\sigma_b = 0$, the $q = 0$ fixed point becomes unstable and the length map acquires a second nonzero fixed point, which is stable. In this bias-free, large weight regime, the network expands small inputs and contracts large inputs. Also, for any nonzero bias b, the length map has a single stable non-zero fixed point. In such a regime, even with small weights, the injected biases at each layer prevent signals from decaying to 0.*

(f) Let's now talk about the stability of fixed points. In a dynamical system, once the system reaches (or starts at) a fixed point, by definition it can never leave. But what happens if the system gets or starts near a fixed point? In real physical systems, this question is very relevant because physical systems almost always have some noise which pushes the system away from a fixed point.

In general, the fixed point can be either stable or unstable. For a stable fixed point, initializing the system near the fixed point will result in behavior which converges to the fixed point, i.e reducing the magnitude of the perturbation away from the fixed point. Conversely, for an unstable fixed point, the system initialized nearby will be repelled from the fixed point.

Use the derivative of the length map at a fixed point to derive conditions on the stability of the fixed point.

Solution If the absolute value of the derivative $\frac{df}{dx}$, evaluated at the fixed point x^*, is less than 1, then the system is stable. This can be seen from considering initializing the system near the fixed point, say at $x^* + \Delta x$. After going through the length map, the value will be

$$\begin{aligned} f(x^* + \Delta x) &\approx f(x^*) + f'(x^*)\Delta x \\ &= x^* + f'(x^*)\Delta x \end{aligned}$$

So the deviation from the fixed point x^* has changed to $f'(x^*)\Delta x$. If the magnitude of $f'(x^*)$ is less than 1, then the magnitude of this deviation is lower than Δx, the system is getting closer to the fixed point, and the fixed point is said to be stable.
Conversely, if the magnitude of $f'(x^*)$ is greater than 1, then the deviations away from equilibrium grow, and the equilibrium is unstable.

(g) With this understanding of stability, revisit your result in part (e) for the tanh nonlinearity.

Specifically, discuss the stability of the fixed points in each of the three regimes. You can estimate the derivative of the length map by looking at the graphs.

Solution
See the italicized paragraph in the solutions above, from the transient chaos paper. In regime (i), there is a single fixed point, $q^* = 0$, and it is stable. In regime (ii), there are two fixed points, $q^* = 0$ (unstable) and some other positive value (stable), and in regime (iii), there is only a positive fixed point, which is stable.

(h) **Do the same stability analysis for the ReLU network.**

Solution In the $\sigma_b = 0$ case, where the only fixed point is at $q = 0$, that point is stable if $\sigma_w < \sqrt{2}$ (because then the slope of the line is less than unity) and unstable if $\sigma_w > \sqrt{2}$. Even for non-zero σ_b, the fixed point (which will now be non-zero) is stable if $\sigma_w < \sqrt{2}$.

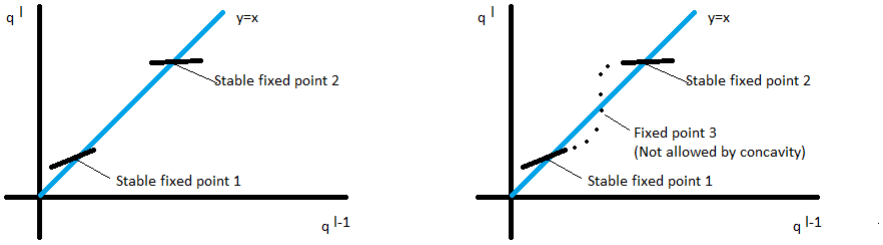
The slightly strange case is when $\sigma_w = \sqrt{2}$ exactly, and $\sigma_b = 0$. In this case, the recurrence relation gives $q^l(q^{l-1}) = q^{l-1}$, meaning that every point is a fixed point. In this case, the fixed points are neither stable nor unstable, since perturbations from them will neither grow or shrink.

(i) **(Optional) You should have found above that the both the ReLU and** tanh **systems never had more than one stable fixed point. Show that this is a consequence of the concavity of the length map.**

Hint: You can just draw a picture for this one. Consider using the fact that the length map is concave, which we discussed in part (c).

Solution

Having two stable fixed points would mean having two intersection points with the line $y = x$ at which the slope of the function is less than unity. But this means that in both cases we approach the function from above, which means that there must have been a third intersection point in the middle. But we already proved that because of the concavity of the length map, the system can have at most two fixed points.



Problem 4: Correlation maps

In the previous problem, we discovered a very interesting property of wide neural networks: the existence of fixed points of activation vector lengths. In this problem, we will explore a similar analysis for correlations between activations due to different inputs to the network and connect this to vanishing/exploding gradients.

Define the correlation q_{12}^l between two inputs $x^{0,1}$ and $x^{0,2}$ at the l^{th} layer of the network via the following inner product:

$$q_{12}^l = \frac{1}{N_L} \sum_{i=1}^{N_l} h_i^l(x^{0,1}) h_i^l(x^{0,2}), \tag{10}$$

This is an important quantity to study, since the effect of the net on inputs that are initially highly correlated (or are not) is relevant to its smoothness.

(a) As in the previous problem, though the quantity q_{12}^l is defined for a specific realization of the network, averaged over all neurons, we can use the self-averaging assumption treat this as an estimate of a probabilistic quantity which characterizes a single neuron but averaged over realizations of the network.

What is this quantity (it is a quantity that comes up quite often when dealing with correlated random variables)?

Solution
The expression on the right hand side of the given equation is an empirical expectation of the quantity

$$\langle h_i^l(x^{0,1}) h_i^l(x^{0,2}) \rangle. \tag{11}$$

Since the random variables $h_i^l(x^{0,1})$ and $h_i^l(x^{0,2})$ are zero-mean, the quantity above is simply the covariance between the neuron's pre-activations corresponding to each of the inputs.

(b) Now we want to find a recurrence relation to describe the behavior of q_{12}^l as we go through the network.

Show that

$$\mathcal{C}\left(c_{12}^{l-1}, q_{11}^{l-1}, q_{22}^{l-1} | \sigma_w, \sigma_b\right) = q_{12}^l \equiv \sigma_w^2 \int_{\mathbb{R}} \mathcal{D}z_1 \mathcal{D}z_2 \phi\left(u_1\right) \phi\left(u_2\right) + \sigma_b^2$$
$$u_1 = \sqrt{q_{11}^{l-1}} z_1, \quad u_2 = \sqrt{q_{22}^{l-1}} \left[c_{12}^{l-1} z_1 + \sqrt{1 - \left(c_{12}^{l-1}\right)^2} z_2 \right],$$

where $\mathcal{D}z_1$ and $\mathcal{D}z_2$ indicate integration with respect to two independent standard normal (Gaussian) variables z_1 and z_2. Here $c_{12}^l \equiv q_{12}^l (q_{11}^l q_{22}^l)^{-1/2}$ is the normalized covariance, also known as the Pearson correlation coefficient.

Hint: Don't be scared by the ugly-looking definitions of u_1 and u_2. This is just done so that we can have the integral over independent Gaussians. Once you have an expression in terms of dependent Gaussians, a simple change-of-variables should get the result you see above.

Solution

$$\begin{aligned}
\mathrm{Cov}(u_1, u_2) = \mathbb{E}(u_1 u_2) &= \mathbb{E}\left(\sqrt{q_{11}^{l-1}} z_1 \cdot \sqrt{q_{22}^{l-1}} \left(c_{12}^{l-1} z_1 + \sqrt{1 - (c_{12}^{l-1})^2} z_2 \right) \right) \\
&= \sqrt{q_{11}^{l-1} q_{22}^{l-1}} \cdot \left(c_{12}^{l-1} \mathbb{E}(z_1^2) + \sqrt{1 - (c_{12}^{l-1})^2} \mathbb{E}(z_1 z_2) \right) \\
&= q_{12}^{l-1} \text{ since } \sqrt{q_{11}^{l-1} q_{22}^{l-1}} \cdot c_{12}^{l-1} = \sqrt{q_{11}^{l-1} q_{22}^{l-1}} \frac{q_{12}^{l-1}}{\sqrt{q_{11}^{l-1} q_{22}^{l-1}}} = q_{12}^{l-1}
\end{aligned}$$

(c) **What are the allowed values of c_{12}^l?**

Solution
We can see from the first definition, given at the beginning of the problem, that $-1 \leq c_{12}^l \leq 1$.

(d) It turns out that the recurrence relation you derived above always has a fixed point at $c^* = 1$.

Show this analytically, and then argue why you could have arrived at this result without any calculation.

Solution
First we show it analytically. If $c^* = 1$, then the variables u_1 and u_2 become

$$u_1 = \sqrt{q^*} z_1$$
$$u_2 = \sqrt{q^*} z_1$$

(there is no z_2 anymore). This gives the expression

$$\sigma_w^2 \int_{\mathbb{R}} \mathcal{D}z_1 \mathcal{D}z_2 \phi\left(z_1\sqrt{q^*}\right)^2 + \sigma_b^2. \tag{12}$$

The random variable z_2 integrates to 1, and so we are left with

$$\sigma_w^2 \int_{\mathbb{R}} \mathcal{D}z_1 \phi\left(z_1\sqrt{q^*}\right)^2 + \sigma_b^2. \tag{13}$$

This is exactly the expression which defines q^*, and since we divide by q^* to get the correlation coefficient, this shows that the correlation coefficient fixed point is unity.
Second, we argue without any calculation. $c^* = 1$ means that the pre-activations corresponding to both inputs are exactly the same. Clearly this cannot be changed by the network, i.e. if we feed it the same input, we get the same output. So perfect correlation is always maintained by the network.

(e) Once you have arrived at the above recurrence relation, note that you have a dynamical system which determines the value of q_{12}^l in terms of q_{12}^{l-1}, q_{11}^{l-1}, and q_{22}^{l-1}. You know from the previous problem, however, that q_{11}^{l-1} and q_{22}^{l-1} converge to fixed points. Though you'll have to take it on faith for now, it turns out that the convergence of these values to their fixed point happens quickly (compared to the dynamics of q_{12}^l), so it is a reasonable approximation to just replace q_{11}^{l-1} and q_{22}^{l-1} with q^*. Then, we have

$$c_{12}^l \approx \frac{1}{q^*}\mathcal{C}\left(c_{12}^{l-1}, q^*, q^* | \sigma_w, \sigma_b\right)$$

Now let's examine the stability of the fixed point $c^* = 1$. To do this, as before, we look at the derivative of the recurrence relation, i.e. dc_{12}^l/dc_{12}^{l-1}, evaluated at the point $c^* = 1$. We will call this quantity χ. It can be shown that

$$\chi = \sigma_w^2 \int_{\mathbb{R}} \mathcal{D}z \left[\phi'(\sqrt{q^*}z)\right]^2 \tag{14}$$

If you want, prove the above relation. Since this is just algebra and not too instructive, feel free to skip this part.

Solution Directly differentiating the integral definition of c_{12}^l with respect to its input, we have

$$\frac{\partial c_{12}^l}{c_{12}^{l-1}} = \sigma_w^2 \int_{\mathbb{R}} \phi'(u_1)\phi'(u_2)\mathcal{D}z_1\mathcal{D}z_2$$

where u_1, u_2 are defined as in part (b). Applying the chain rule and then integrating by parts, we can simplify this integral to

$$\chi \equiv \frac{\partial c_{12}^l}{c_{12}^{l-1}} = \sigma_w^2 \int_{\mathbb{R}} \mathcal{D}z \left[\phi'(\sqrt{q^*}z)\right]^2$$

as desired.

Problem 5: Connection between χ and exploding/vanishing gradients

In the previous problem, we saw that a unit correlation coefficient (the highest value which is allowed) is a fixed point under evolution through the neural network. However, it was not always stable. The stability depended on whether the quantity χ, defined as the derivative of dc_{12}^l/dc_{12}^{l-1}, is greater or less than 1. In this problem, we want to understand the connection between the quantity χ and exploding or vanishing gradients.

A starting point to understand the connection between χ and gradients is to consider what it means for the fixed point $c^* = 1$ to be stable or unstable. If $\chi > 1$, and the fixed point $c^* = 1$ is thus unstable, then two input vectors which are highly correlated will de-correlate as they are processed by the network. Conversely, if $\chi < 1$, then two vectors will become more correlated as they are processed. Thus it seems like if $\chi > 1$, space is stretched, while if $\chi < 1$, space is contracted. In this problem we find that χ is precisely the factor by which space is streched or contracted by each layer of the neural network.

(a) Start by considering a plain linear transformation, $\mathbf{y} = \mathbf{J}\mathbf{x}$, where $\mathbf{x}$ and $\mathbf{y}$ are vectors and $\mathbf{J}$ is a matrix.

Averaged over all possible directions in which x can point, what is the ratio of the squared length of y to that of x, in terms of the singular values of J?

Solution
The length of y is given by the square root of $\mathbf{y}^T\mathbf{y}$, where

$$\mathbf{y}^T\mathbf{y} = \mathbf{x}^T\mathbf{J}^T\mathbf{J}\mathbf{x} \tag{15}$$

Even though the matrix $\mathbf{J}$ might not admit an eigendecomposition, the matrix $\mathbf{J}^T\mathbf{J}$ does, so we can express it as

$$\mathbf{J}^T\mathbf{J} = \sum_i \lambda_i \mathbf{v}_i \mathbf{v}_i^T, \tag{16}$$

where λ_i and $\mathbf{v}_i$ are the i^{th} eigenvalue and eigenvector of $\mathbf{J}^T\mathbf{J}$. Then

$$\mathbf{y}^T\mathbf{y} = \sum_i \lambda_i (\mathbf{x}^T\mathbf{v}_i)^2 \tag{17}$$

Without loss of generality we can consider $\mathbf{x}$ to have unit length and calculate the expectation of the squared length of $\mathbf{y}$. Averaged over all the possible directions of $\mathbf{x}$:

$$\langle \mathbf{y}^T\mathbf{y} \rangle = \sum_i \lambda_i \langle (\mathbf{x}^T\mathbf{v}_i)^2 \rangle \tag{18}$$

The squared projection of $\mathbf{x}$ onto a vector $\mathbf{v}_i$ will average to $1/N$, where N is the dimension of the vector. So we find that $\langle \mathbf{y}^T\mathbf{y} \rangle$ is given by the mean eigenvalue of $\mathbf{J}^T\mathbf{J}$. Since the eigenvalues of $\mathbf{J}^T\mathbf{J}$ are squared singular values of $\mathbf{J}$, the squared length increases by the mean squared singular value of $\mathbf{J}$.

(b) Consider the transformation enacted by layer l of a neural network,

$$h^l = W^l \phi(h^{l-1}) + b^l. \tag{19}$$

Let D be a diagonal matrix whose entries are $D_{ii} = \phi'(h_i^{l-1})$.

In terms of W^l and D, what is the Jacobian of the transformation $h^{l-1} \mapsto h^l$?

Solution
Since $h_i^l = \sum_j W_{ij}^l \phi(h_j^{l-1})$,

$$\frac{\partial h_i^l}{\partial h_j^{l-1}} = W_{ij}^l \phi'(h_j^{l-1}). \tag{20}$$

Thus the Jacobian is simply given by the product of W^l and D.

From the problem above, we want the mean-squared singular value of the Jacobian. Show that this mean-squared singluar value is exactly equal to the expression for χ calculated earlier, when we take the expectation with respect to the distribution of the weight matrix and the distribution of the pre-activations.

Solution The singular values are the eigenvalues of $\mathbf{J}^T\mathbf{J}$, so the mean squared singular value is $\frac{1}{N}\text{tr}\left(\mathbf{J}^T\mathbf{J}\right) = \frac{1}{N}\text{tr}(D^l W^l)$. Of course, since W and D are random matrices, this is a random variable; we want $\mathbb{E}\left(\frac{1}{N}\text{tr}(D^l W^l)\right)$. D is a diagonal matrix with entries $\phi'(h_i^{l-1})$; as N becomes very large and we make the mean-field approximation wherein the weighted combination of the entries of D by i.i.d. random Gaussian variables becomes integration against the Gaussian measure, we have

$$\chi = \sigma_w^2 \int_{\mathbb{R}} \phi'\left(\sqrt{q^*}\rho\right) \mathcal{D}\rho$$

as desired (where $\mathcal{D}\rho$ is the standard Gaussian measure).

(c) How does the value of χ relate to exploding and vanishing gradients?

Solution
If $\chi > 1$, we saw from the above problem that perturbations get magnified as they go through a layer, and if $\chi < 1$, we saw that perturbations get shrunk as they go through a layer. So the regime in which $\chi > 1$ corresponds to the regime of exploding gradients, while the regime in which $\chi < 1$ corresponds to vanishing gradients.

(d) To emphasize that χ is a function of σ_w and σ_b, we go back to the expression for χ,

$$\chi = \sigma_w^2 \int_{\mathbb{R}} \mathcal{D}z\phi'(\sqrt{q^*}z)^2, \tag{21}$$

and explicitly write the dependence of the fixed point q^* on σ_w and σ_b. Remember that this dependence can be written implicitly via the equation

$$q^* = \sigma_w^2 \int_{\mathbb{R}} \phi\left(\rho\sqrt{q^*}\right)^2 \mathrm{d}\rho + \sigma_b^2. \tag{22}$$

Numerically calculate the value of χ for several values of σ_w and σ_b, for the hyperbolic tangent nonlinearity. Make a contour plot, and specifically indicate the curve corresponding to $\chi = 1$

Solution
The following Python code should work:

```
import numpy as np
import scipy.integrate as integrate
import scipy.optimize as optimize

def gaussian(x):
    return np.sqrt(2 * np.pi) * np.exp(-0.5 * x**2)

def integrand(x):
    return np.tanh(x * np.sqrt(q))**2 * gaussian(x)

def q_star(starting_q, sigma_w, sigma_b):
    def integral(q):
        return sigma_w**2 * integrate.quad(integrand, -np.inf, np.inf)[0] + sigma
    return optimize.fixed_point(integral, starting_q)

def chi(sigma_w, sigma_b):
    q = q_star(0.7, sigma_w, sigma_b)
    # Derivative of hyperbolic tangent is hyperbolic secant.
    return sigma_w**2 * integrate.quad(
        lambda x: 1 / np.cosh(x * np.sqrt(q))**2 * gaussian(x),
        -np.inf, np.inf)
```