CS 159—HW #3

Due 29 April 2021

Instructions

- The homework is scored out of 20 points. Question 5 requires coding and is worth 9 points.
- Complete the homework in groups of two to three, clearly listing the group members at the top of the solution. Submit only one solution per group on Gradescope.
- Either LaTeX or handwritten solutions are fine—just make sure the solution is legible. In either case turn in a pdf of your solution.

1 MLP sketch (2 points)

Consider a multilayer perceptron (MLP) defined recursively by the following equations:

$$z_i^1 = \sum_{j=1}^{d_0} W_{ij}^1 x_j, \tag{1}$$

$$z_i^l = \sum_{j=1}^{d_{l-1}} W_{ij}^l \varphi\left(z_j^{l-1}\right), \tag{2}$$

where x_j refers to the jth component of the input, z_i^l refers to the ith component of the lth layer's pre-activations, W^l refers to the weight matrix at the lth layer, and φ is the nonlinearity.

Consider such a network with 4 layers with input dimension $d_0 = 3$, hidden dimensions $d_1 = d_2 = d_3 = 4$ and output dimension $d_4 = 1$. Draw a diagram of this network and label the weight matrices and hidden layers.

2 Wiring constraints (3 points)

Consider a "linear neuron" $y = \sum_{i=1}^d w_i x_i$ where w_i are the weights and x_i are the inputs. Suppose that the weights satisfy two constraints:

1. $\sum_{i=1}^d w_i = 0$;
2. $\sum_{i=1}^d w_i^2 = 1$.

Further, suppose that the input components $x_1, x_2, ..., x_d$ are uncorrelated random variables each with mean μ and variance σ^2. Compute $\mathbb{E}[y]$ and $\mathrm{Var}[y]$, and interpret the results.

Hint: when is the variance of a sum equal to the sum of the variances?

3 Deep perturbation theory (3 points)

Consider the 2-layer "deep linear network" defined by:

$$y(x; W_2, W_1) := W_2 W_1 x,$$

where the input $x \in \mathbb{R}^{d_0}$ and neither of the matrices $W_2 \in \mathbb{R}^{d_2 \times d_1}$ or $W_1 \in \mathbb{R}^{d_1 \times d_0}$ is zero.

Show that for any input x and for any two perturbation matrices $\Delta W_2 \in \mathbb{R}^{d_2 \times d_1}$ and $\Delta W_1 \in \mathbb{R}^{d_1 \times d_0}$, this network satisfies the following perturbation bound:

$$\frac{\|y(x; W_2 + \Delta W_2, W_1 + \Delta W_1) - y(x; W_2, W_1)\|_2}{\|y(x; W_2, W_1)\|_2} \leq C \left[\left(1 + \frac{\|\Delta W_2\|_F}{\|W_2\|_F}\right) \left(1 + \frac{\|\Delta W_1\|_F}{\|W_1\|_F}\right) - 1 \right]$$

where $\|\cdot\|_2$ denotes the vector ℓ_2 norm, $\|\cdot\|_F$ denotes the Frobenius norm, and $C \geq 0$ is a constant (involving spectral properties of matrices in the problem) that you need not specify.

Hint: start by multiplying out the righthand side. Next multiply out the lefthand side and try to bring it into a form matching the righthand side by using the following bounds. For the numerator, you'll need the triangle inequality and the fact that $\|AB\|_F \leq \|A\|_F \|B\|_F$. For the denominator, you'll need the fact that $\|AB\|_F \geq c_A \|A\|_F \|B\|_F$, where $c_A \geq 0$ depends on spectral properties of A that you need not work out.

Guess the form of a similar result for the 3-layer deep linear network.

4 Linear neuron yields a Gaussian process (3 points)

Consider the linear neuron $y(x) = \sum_{i=1}^{d} w_i x_i$. For a collection of n input vectors $x^{(1)}, x^{(2)}, ..., x^{(n)}$, show that if the weights $\{w_i\}_{i=1}^{d}$ are drawn iid $\mathcal{N}(0, \sigma^2)$, then the outputs $y(x^{(1)}), y(x^{(2)}), ..., y(x^{(n)})$ are jointly Normal. Conclude that the linear neuron yields a Gaussian process.

Hint: a linear transformation of a Gaussian random vector is still Gaussian.

Compute the mean $\mathbb{E}[y(x^{(1)})]$ and second moments $\mathbb{E}[y(x^{(1)})y(x^{(2)})]$ of this Gaussian process.

5 Verifying the relu MLP covariance (9 points)

Consider an L-layer MLP (as defined in Question 1) with a 1-dimensional output $z^L \in \mathbb{R}$. For four inputs $x^{(1)}, x^{(2)}, x^{(3)}, x^{(4)} \in \mathbb{R}^{d_0}$ and a particular setting of the weights, the network has four outputs $z^L(x^{(1)}), z^L(x^{(2)}), z^L(x^{(3)}), z^L(x^{(4)}) \in \mathbb{R}$.

In lecture 8, we considered the special case that:

1. the nonlinearity is set to $\varphi(z) = \sqrt{2} \cdot \max(0, z)$;
2. each input is scaled to have Euclidean norm $\|x^{(i)}\|_2 = \sqrt{d_0}$;
3. the weights in the lth layer are drawn iid from $\mathcal{N}\left(0, \frac{1}{d_l}\right)$, where d_l is the width of the layer.

Under these conditions, we claimed that if the hidden layer widths $d_1, d_2, ..., d_{L-1}$ are taken to infinity, then the distribution of outputs is multivariate Normal with mean zero and covariance:

$$\mathbb{E}\left[z^L(x^{(i)}) z^L(x^{(j)})\right] = \underbrace{h \circ ... \circ h}_{L-1 \text{ times}} \left(\frac{x^{(i)T} x^{(j)}}{d_0}\right).$$

In this expression, $\circ$ denotes function composition and the function h is defined by:

$$h(t) := \tfrac{1}{\pi}\left[\sqrt{1-t^2} + t \cdot (\pi - \arccos t)\right].$$

Your task is to verify this expression for the covariance empirically. We have provided a Jupyter notebook (q5.ipynb) to help you do this. The notebook already takes care of much of the work: it constructs four properly scaled inputs, constructs a network with properly scaled nonlinearities and has a loop that samples random networks with properly scaled weights. To avoid needing to install Python dependencies, note that the notebook should run natively on `https://colab.research.google.com/`—set the runtime to GPU.

Extend this notebook by completing the last two cells to compute both the empirical output covariance for 1000 random networks, and the theoretical covariance using the infinite width result. Turn in the two covariance matrices (i.e. two 4×4 matrices).