Topics in Artificial Intelligence (CPSC 532S): Assignment #1

Due on Monday, January 21, 2021 at 11:59pm PST

In this assignment you will get hands on experience with the basic operations, processing and inner-workings of traditional deep learning libraries. There are many deep learning libraries that allow easy creation of complex neural network architectures. In future assignments, for example, we will use **PyTorch** to do this (other popular ones are Keras, TensorFlow and Theano). However, those libraries in an interest of ease of use often abstract a lot of detail. The basics learned in this assignment will give you hands on experience on how they work under the hood and give you skills necessary to implement new types of layers and whole architectures (if necessary) and to think through corresponding algorithmic and computational issues.

Problem 1 (40 points)

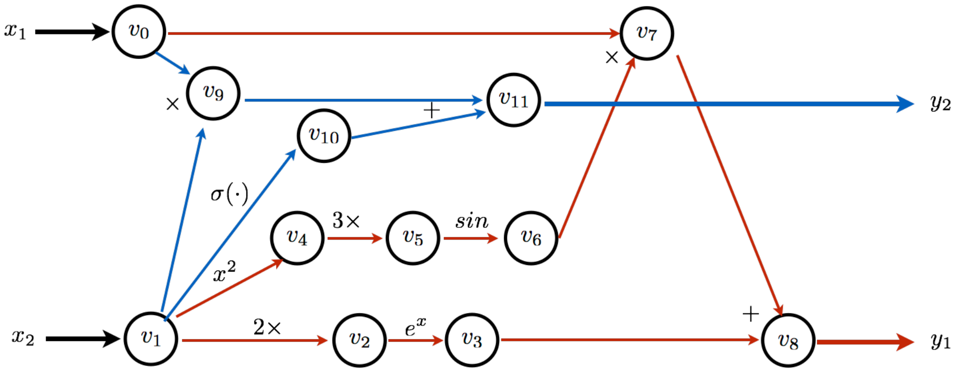
The key to learning in deep neural networks is ability to compute the derivative of a vector function $\mathbf{f}$ with respect to the parameters of the deep neural network. Computing such derivatives as closed-form expressions for complex functions $\mathbf{f}$ is difficult and computationally expensive. Therefore, instead, deep learning packages define computational graphs and use automatic differentiation algorithms (typically backpropagation) to compute gradients progressively through the network using the chain rule.

Let us define the following vector functions:

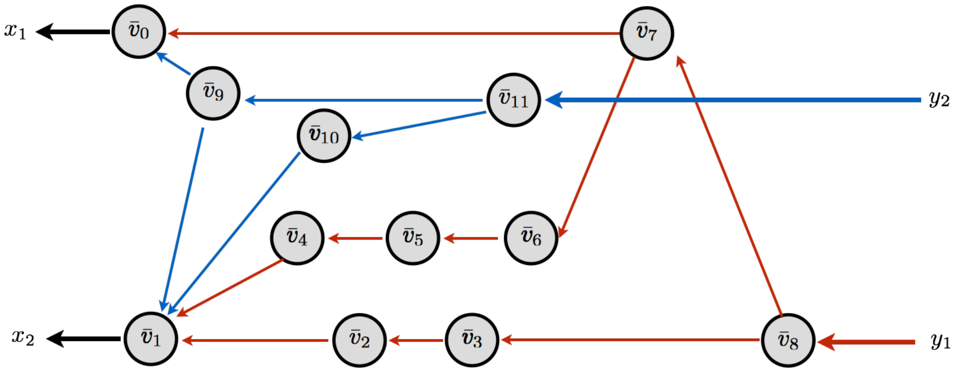
$$\begin{aligned} y_1 &= \mathbf{f}_1(x_1, x_2) = e^{2x_2} + x_1 sin(3x_2^2) && (1) \\ y_2 &= \mathbf{f}_2(x_1, x_2) = x_1 x_2 + \sigma(x_2), && (2) \end{aligned}$$

where $\sigma(\cdot)$ denotes the standard sigmoid function. This is equivalent to a network with two inputs $\mathbf{x} = \begin{bmatrix} x_1 \\ x_2 \end{bmatrix}$ and two outputs $\mathbf{y} = \begin{bmatrix} y_1 \\ y_2 \end{bmatrix} = \mathbf{f}(\mathbf{x})$ and a set of intermediate layers (note that bold indicates vectors).

(a) Draw the computational graph.



(b) Draw backpropagation graph.



(c) Compute the value of $\mathbf{f}(\mathbf{x})$ at $\mathbf{x} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$.

This requires only one pass irrespective of number of inputs or outputs:

Expression	**Evaluation**
$v_0 = x_1$	2
$v_1 = x_2$	1
$v_2 = 2v_1$	$2 \times 1 = 2$
$v_3 = e^{v_2}$	$e^2 = 7.389$
$v_4 = (v_1)^2$	$1^2 = 1$
$v_5 = 3v_4$	$3 \times 1 = 3$
$v_6 = sin(v_5)$	$sin(3) = 0.141$
$v_7 = v_0 v_6$	$2 \times 0.141 = 0.282$
$v_8 = v_7 + v_3$	$0.282 + 7.389 = 7.671$
$v_9 = v_0 v_1$	$1 \times 2 = 2$
$v_{10} = \sigma(v_1)$	$\sigma(1) = \frac{1}{1+e^{-1}} = 0.731$
$v_{11} = v_9 + v_{10}$	$2 + 0.731 = 2.731$

Answer: $\mathbf{f}\left(\begin{bmatrix} 2 \\ 1 \end{bmatrix}\right) = \begin{bmatrix} v_8 \\ v_{11} \end{bmatrix} = \begin{bmatrix} 7.671 \\ 2.731 \end{bmatrix}$

(d) At $\mathbf{x} = \begin{pmatrix} 2 \\ 1 \end{pmatrix}$, compute Jacobian using forward mode auto-differentiation.

Forward mode auto-differentiation requires one pass for each *input*.

For computing $\frac{\partial y_1}{\partial x_1}$ and $\frac{\partial y_2}{\partial x_1}$:

Expression	**Rule**	**Evaluation**
$\frac{\partial v_0}{\partial x_1}$	-	1
$\frac{\partial v_1}{\partial x_1}$	-	0
$\frac{\partial v_2}{\partial x_1} = \frac{\partial v_2}{\partial v_1}\frac{\partial v_1}{\partial x_1}$	Chain Rule	$2 \times 0 = 0$
$\frac{\partial v_3}{\partial x_1} = \frac{\partial v_3}{\partial v_2}\frac{\partial v_2}{\partial x_1}$	Chain Rule	$e^{v_2} \times 0 = 0$
$\frac{\partial v_4}{\partial x_1} = \frac{\partial v_4}{\partial v_1}\frac{\partial v_1}{\partial x_1}$	Chain Rule	$2v_1 \times 0 = 2 \times 0 = 0$
$\frac{\partial v_5}{\partial x_1} = \frac{\partial v_5}{\partial v_4}\frac{\partial v_4}{\partial x_1}$	Chain Rule	$3 \times 0 = 0$
$\frac{\partial v_6}{\partial x_1} = \frac{\partial v_6}{\partial v_5}\frac{\partial v_5}{\partial x_1}$	Chain Rule	$cos(v_5) \times 0 = 0$
$\frac{\partial v_7}{\partial x_1} = \frac{\partial v_7}{\partial v_6}\frac{\partial v_6}{\partial x_1} + \frac{\partial v_7}{\partial v_0}\frac{\partial v_0}{\partial x_1}$	Product Rule	$v_0 \times 0 + v_6 \times 1 = 0.141$
$\frac{\partial v_8}{\partial x_1} = \frac{\partial v_7}{\partial x_1} + \frac{\partial v_3}{\partial x_1}$	Sum Rule	$0.141 + 0 = 0.141$
$\frac{\partial v_9}{\partial x_1} = \frac{\partial v_9}{\partial v_0}\frac{\partial v_0}{\partial x_1} + \frac{\partial v_9}{\partial v_1}\frac{\partial v_1}{\partial x_1}$	Product Rule	$v_1 \times 1 + v_0 \times 0 = 1$
$\frac{\partial v_{10}}{\partial x_1} = \frac{\partial v_{10}}{\partial v_1}\frac{\partial v_1}{\partial x_1}$	Chain Rule	$\sigma(v_1)(1 - \sigma(v_1)) \times 0 = 0$
$\frac{\partial v_{11}}{\partial x_1} = \frac{\partial v_9}{\partial x_1} + \frac{\partial v_{10}}{\partial x_1}$	Sum Rule	$1 + 0 = 1$
$\frac{\partial y_1}{\partial x_1} = \frac{\partial v_8}{\partial x_1}$	-	0.141
$\frac{\partial y_2}{\partial x_1} = \frac{\partial v_{11}}{\partial x_1}$	-	1

For computing $\frac{\partial y_1}{\partial x_2}$ and $\frac{\partial y_2}{\partial x_2}$:

Expression	**Rule**	**Evaluation**
$\frac{\partial v_0}{\partial x_2}$	-	0
$\frac{\partial v_1}{\partial x_2}$	-	1
$\frac{\partial v_2}{\partial x_2} = \frac{\partial v_2}{\partial v_1}\frac{\partial v_1}{\partial x_2}$	Chain Rule	$2 \times 1 = 2$
$\frac{\partial v_3}{\partial x_2} = \frac{\partial v_3}{\partial v_2}\frac{\partial v_2}{\partial x_2}$	Chain Rule	$e^{v_2} \times 2 = 14.778$
$\frac{\partial v_4}{\partial x_2} = \frac{\partial v_4}{\partial v_1}\frac{\partial v_1}{\partial x_2}$	Chain Rule	$2v_1 \times 1 = 2 \times 1 = 2$
$\frac{\partial v_5}{\partial x_2} = \frac{\partial v_5}{\partial v_4}\frac{\partial v_4}{\partial x_2}$	Chain Rule	$3 \times 2 = 6$
$\frac{\partial v_6}{\partial x_2} = \frac{\partial v_6}{\partial v_5}\frac{\partial v_5}{\partial x_2}$	Chain Rule	$cos(v_5) \times 6 = ?5.940$
$\frac{\partial v_7}{\partial x_2} = \frac{\partial v_7}{\partial v_6}\frac{\partial v_6}{\partial x_2} + \frac{\partial v_7}{\partial v_0}\frac{\partial v_0}{\partial x_2}$	Product Rule	$v_0 \times ?5.940 + v_6 \times 0 = ?11.880$
$\frac{\partial v_8}{\partial x_2} = \frac{\partial v_7}{\partial x_2} + \frac{\partial v_3}{\partial x_2}$	Sum Rule	$?11.880 + 14.778 = 2.898$
$\frac{\partial v_9}{\partial x_2} = \frac{\partial v_9}{\partial v_0}\frac{\partial v_0}{\partial x_2} + \frac{\partial v_9}{\partial v_1}\frac{\partial v_1}{\partial x_2}$	Product Rule	$v_1 \times 0 + v_0 \times 1 = 2$
$\frac{\partial v_{10}}{\partial x_2} = \frac{\partial v_{10}}{\partial v_1}\frac{\partial v_1}{\partial x_2}$	Chain Rule	$\sigma(v_1)(1 - \sigma(v_1)) \times 1 = 0.197$
$\frac{\partial v_{11}}{\partial x_2} = \frac{\partial v_9}{\partial x_2} + \frac{\partial v_{10}}{\partial x_2}$	Sum Rule	$2 + 0.197 = 2.197$
$\frac{\partial y_1}{\partial x_2} = \frac{\partial v_8}{\partial x_2}$	-	2.898
$\frac{\partial y_2}{\partial x_2} = \frac{\partial v_{11}}{\partial x_2}$	-	2.197

Answer: **Jacobian** $= \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 0.141 & 2.898 \\ 1 & 2.197 \end{bmatrix}$

(e) At $\mathbf{x} = \begin{pmatrix}2\\1\end{pmatrix}$, compute Jacobian using backward mode auto-differentiation.

Backward mode auto-differentiation requires one pass for each *output*. The backward mode also requires working with the adjoint nodes and graph.

For computing $\frac{\partial y_1}{\partial x_1}$ and $\frac{\partial y_1}{\partial x_2}$:

Expression	**Evaluation**
$\bar{v}_{11} = \frac{\partial y_1}{\partial v_{11}}$	0
$\bar{v}_{10} = \bar{v}_{11}\frac{\partial v_{11}}{\partial v_{10}}$	$0 \times 1 = 0$
$\bar{v}_9 = \bar{v}_{11}\frac{\partial v_{11}}{\partial v_9}$	$0 \times 1 = 0$
$\bar{v}_8 = \frac{\partial y_1}{\partial v_8}$	1
$\bar{v}_7 = \bar{v}_8\frac{\partial v_8}{\partial v_7}$	$1 \times 1 = 0$
$\bar{v}_6 = \bar{v}_7\frac{\partial v_7}{\partial v_6}$	$1 \times v_0 = 2$
$\bar{v}_5 = \bar{v}_6\frac{\partial v_6}{\partial v_5}$	$2 \times cos(v_5) = -1.980$
$\bar{v}_4 = \bar{v}_5\frac{\partial v_5}{\partial v_4}$	$-1.980 \times 3 = -5.940$
$\bar{v}_3 = \bar{v}_8\frac{\partial v_8}{\partial v_3}$	$1 \times 1 = 1$
$\bar{v}_2 = \bar{v}_3\frac{\partial v_3}{\partial v_2}$	$1 \times e^{v_2} = 7.389$
$\bar{v}_1 = \bar{v}_2\frac{\partial v_2}{\partial v_1} + \bar{v}_4\frac{\partial v_4}{\partial v_1} + \bar{v}_9\frac{\partial v_9}{\partial v_1} + \bar{v}_{10}\frac{\partial v_{10}}{\partial v_1}$	$[7.389 \times 2] + [-5.940 \times 2v_1] +$
	$+[0 \times v_0] + [0 \times \sigma(v_1)(1-\sigma(v_1))] = 2.898$
$\bar{v}_0 = \bar{v}_7\frac{\partial v_7}{\partial v_0} + \bar{v}_9\frac{\partial v_9}{\partial v_0}$	$[1 \times v_6] + [0 \times v_1] = 0.141$

For computing $\frac{\partial y_2}{\partial x_1}$ and $\frac{\partial y_2}{\partial x_2}$:

Expression	**Evaluation**
$\bar{v}_{11} = \frac{\partial y_1}{\partial v_{11}}$	1
$\bar{v}_{10} = \bar{v}_{11}\frac{\partial v_{11}}{\partial v_{10}}$	$1 \times 1 = 1$
$\bar{v}_9 = \bar{v}_{11}\frac{\partial v_{11}}{\partial v_9}$	$1 \times 1 = 1$
$\bar{v}_8 = \frac{\partial y_1}{\partial v_8}$	0
$\bar{v}_7 = \bar{v}_8\frac{\partial v_8}{\partial v_7}$	$0 \times 1 = 0$
$\bar{v}_6 = \bar{v}_7\frac{\partial v_7}{\partial v_6}$	$0 \times v_0 = 0$
$\bar{v}_5 = \bar{v}_6\frac{\partial v_6}{\partial v_5}$	$0 \times cos(v_5) = 0$
$\bar{v}_4 = \bar{v}_5\frac{\partial v_5}{\partial v_4}$	$0 \times 3 = 0$
$\bar{v}_3 = \bar{v}_8\frac{\partial v_8}{\partial v_3}$	$0 \times 1 = 0$
$\bar{v}_2 = \bar{v}_3\frac{\partial v_3}{\partial v_2}$	$0 \times e^{v_2} = 0$
$\bar{v}_1 = \bar{v}_2\frac{\partial v_2}{\partial v_1} + \bar{v}_4\frac{\partial v_4}{\partial v_1} + \bar{v}_9\frac{\partial v_9}{\partial v_1} + \bar{v}_{10}\frac{\partial v_{10}}{\partial v_1}$	$[0 \times 2] + [0 \times 2v_1] +$
	$+[1 \times v_0] + [1 \times \sigma(v_1)(1-\sigma(v_1))] = 2.197$
$\bar{v}_0 = \bar{v}_7\frac{\partial v_7}{\partial v_0} + \bar{v}_9\frac{\partial v_9}{\partial v_0}$	$[0 \times v_6] + [1 \times v_1] = 1$

Answer: **Jacobian** $= \frac{\partial \mathbf{f}(\mathbf{x})}{\partial \mathbf{x}} = \frac{\partial \mathbf{y}}{\partial \mathbf{x}} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 0.141 & 2.898 \\ 1 & 2.197 \end{bmatrix}$

Problem 2 (15 points)

Consider a neural network with N input units denoted by $\mathbf{x} \in \mathbb{R}^N$ vector, M output units denoted by $\mathbf{y} \in \mathbb{R}^M$ vector, and K hidden units denoted by $\mathbf{h} \in \mathbb{R}^K$ vector. The hidden layer has an activation function σ (e.g., a sigmoid or ReLU). The resulting equations that govern the behavior of this simple network are:

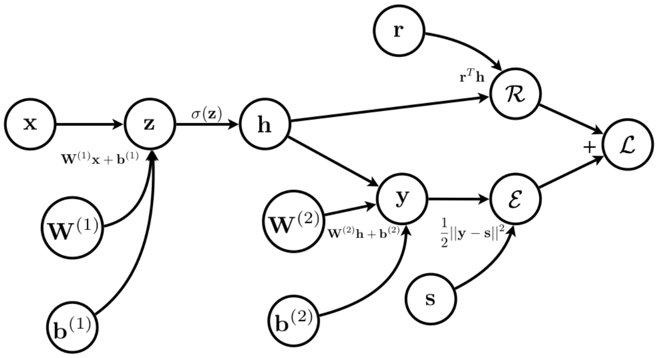
$$\begin{aligned}\mathbf{z} &= \mathbf{W}^{(1)}\mathbf{x} + \mathbf{b}^{(1)} \\ \mathbf{h} &= \sigma(\mathbf{z}) \\ \mathbf{y} &= \mathbf{W}^{(2)}\mathbf{h} + \mathbf{b}^{(2)}\end{aligned}$$

The loss function $\mathcal{L}$ involves an L2 error on prediction $\mathcal{E}$ plus a regularizer $\mathcal{R}$ as follows:

$$\begin{aligned}\mathcal{L} &= \mathcal{E} + \mathcal{R} \\ \mathcal{R} &= \mathbf{r}^T\mathbf{h} \\ \mathcal{E} &= \frac{1}{2}||\mathbf{y} - \mathbf{s}||^2\end{aligned}$$

where $\mathbf{r}$ and $\mathbf{s}$ are given (e.g., $\mathbf{s}$ is a ground truth prediction for $\mathbf{x}$).

(a) Draw the computational graph relating $\mathbf{x}$, $\mathbf{z}$, $\mathbf{h}$, $\mathbf{y}$, $\mathcal{L}$, $\mathcal{E}$, $\mathcal{R}$. Note, in this case we expect nodes to be representing collections of values, not each neuron individually.



(b) Derive the backpropagation equations for computing $\frac{\partial \mathcal{L}}{\partial \mathbf{x}}$. Please use σ' to denote the derivative of the activation function.

$$\hat{\mathcal{R}} = \frac{\partial \mathcal{L}}{\partial \mathcal{R}} = 1$$

$$\hat{\mathcal{E}} = \frac{\partial \mathcal{L}}{\partial \mathcal{E}} = 1$$

$$\hat{\mathbf{y}} = \frac{\partial \mathcal{E}}{\partial \mathbf{y}}\hat{\mathcal{E}} = 1 \times \begin{bmatrix} \frac{\partial \mathcal{E}}{\partial \mathbf{y}_1} \\ \vdots \\ \frac{\partial \mathcal{E}}{\partial \mathbf{y}_n} \end{bmatrix} = \begin{bmatrix} \mathbf{y}_1 - \mathbf{s}_1 \\ \vdots \\ \mathbf{y}_n - \mathbf{s}_n \end{bmatrix} = \mathbf{y} - \mathbf{s}$$

$$\hat{\mathbf{h}} = \frac{\partial \mathcal{R}}{\partial \mathbf{h}}\hat{\mathcal{R}} + \frac{\partial \mathbf{y}}{\partial \mathbf{h}}\hat{\mathbf{y}} = \mathbf{r} \times 1 + \left[\mathbf{W}^{(2)}\right]^T \times (\mathbf{y} - \mathbf{s}) = \mathbf{r} + \left[\mathbf{W}^{(2)}\right]^T (\mathbf{y} - \mathbf{s})$$

Problem 2 (15 points) continued on next page. . .

$$\hat{\mathbf{z}} = \frac{\partial \mathbf{h}}{\partial \mathbf{z}} \hat{\mathbf{h}} = \mathrm{diag}(\sigma'(\mathbf{z})) \times \hat{\mathbf{h}}$$

$$\hat{\mathbf{x}} = \frac{\partial \mathbf{z}}{\partial \mathbf{x}} \hat{\mathbf{z}} = \left[\mathbf{W}^{(1)}\right]^T \hat{\mathbf{z}}$$

Credit: Problem 2 is adopted from University of Toronto's CSC421/2516 Problem Set 1.