Homework 7: Computation Graphs, Back-propagation, and Neural Networks

Due: Friday, May 6th, 2022 at 11:59PM EST

Instructions: Your answers to the questions below, including plots and mathematical work, should be submitted as a single PDF file. It's preferred that you write your answers using software that typesets mathematics (e.g.LaTeX, LyX, or MathJax via iPython), though if you need to you may scan handwritten work. You may find the minted package convenient for including source code in your LaTeX document. If you are using LyX, then the listings package tends to work better.

---

1 Introduction

There is no doubt that neural networks are a very important class of machine learning models. Given the sheer number of people who are achieving impressive results with neural networks, one might think that it's relatively easy to get them working. This is a partly an illusion. One reason so many people have success is that, thanks to GitHub, they can copy the exact settings that others have used to achieve success. In fact, in most cases they can start with "pre-trained" models that already work for a similar problem, and "fine-tune" them for their own purposes. It's far easier to tweak and improve a working system than to get one working from scratch. If you create a new model, you're kind of on your own to figure out how to get it working: there's not much theory to guide you and the rules of thumb do not always work. Understanding even the most basic questions, such as the preferred variant of SGD to use for optimization, is still a very active area of research.

One thing is clear, however: If you do need to start from scratch, or debug a neural network model that doesn't seem to be learning, it can be immensely helpful to understand the low-level details of how your neural network works – specifically, back-propagation. With this assignment, you'll have the opportunity to linger on these low-level implementation details. Every major neural network type (RNNs, CNNs, Resnets, etc.) can be implemented using the basic framework we'll develop in this assignment.

To help things along, Philipp Meerkamp, Pierre Garapon, and David Rosenberg have designed a minimalist framework for computation graphs and put together some support code. The intent is for you to read, or at least skim, every line of code provided, so that you'll know you understand all the crucial components and could, in theory, create your own from scratch. In fact, creating your own computation graph framework from scratch is highly encouraged – you'll learn a lot.

2 Computation Graph Framework

To get started, please read the tutorial on the computation graph framework we'll be working with. (Note that it renders better if you view it locally.) The use of computation graphs is not specific to machine learning or neural networks. Computation graphs are just a way to represent a function that facilitates efficient computation of the function's values and its gradients with respect to inputs. The tutorial takes this perspective, and there is very little in it about machine learning, per se.

To see how the framework can be used for machine learning tasks, we've provided a full implementation of linear regression. You should start by working your way through the `__init__` of the `LinearRegression` class in `linear_regression.py`. From there, you'll want to review the node class definitions in `nodes.py`, and finally the class `ComputationGraphFunction` in `graph.py`. `ComputationGraphFunction` is where we repackage a raw computation graph into something that's more friendly to work with for machine learning. The rest of `linear_regression.py` is fairly routine, but it illustrates how to interact with the `ComputationGraphFunction`.

As we've noted earlier in the course, getting gradient calculations correct can be difficult. To help things along, we've provided two functions that can be used to test the backward method of a node and the overall gradient calculation of a `ComputationGraphFunction`. The functions are in `test_utils.py`, and it's recommended that you review the tests provided for the linear regression implementation in `linear_regression.t.py`. (You can run these tests from the command line with `python3 linear_regression.t.py.`) The functions actually doing the testing, `test_node_backward` and `test_ComputationGraphFunction`, may seem a bit intricate, but they're implementing the exact same `gradient_checker` logic we saw in the second homework assignment.

Once you've understood how linear regression works in our framework, you're ready to start implementing your own algorithms. To help you get started, please make sure you are able to execute the following commands:

- cd /path/to/hw7
- python3 linear_regression.py
- python3 linear_regression.t.py

3 Ridge Regression

When moving to a new system, it's always good to start with something familiar. But that's not the only reason we're doing ridge regression in this homework. In ridge regression the parameter vector is "shared", in the sense that it's used twice in the objective function. In the computation graph, this can be seen in the fact that the node for the parameter vector has two outgoing edges. This sharing is common many popular neural networks (RNNs and CNNs), where it is often referred to as *parameter tying*.

`ridge_regression.py` provides a skeleton code and `ridge_regression.t.py` is a test code, which you should eventually be able to pass.

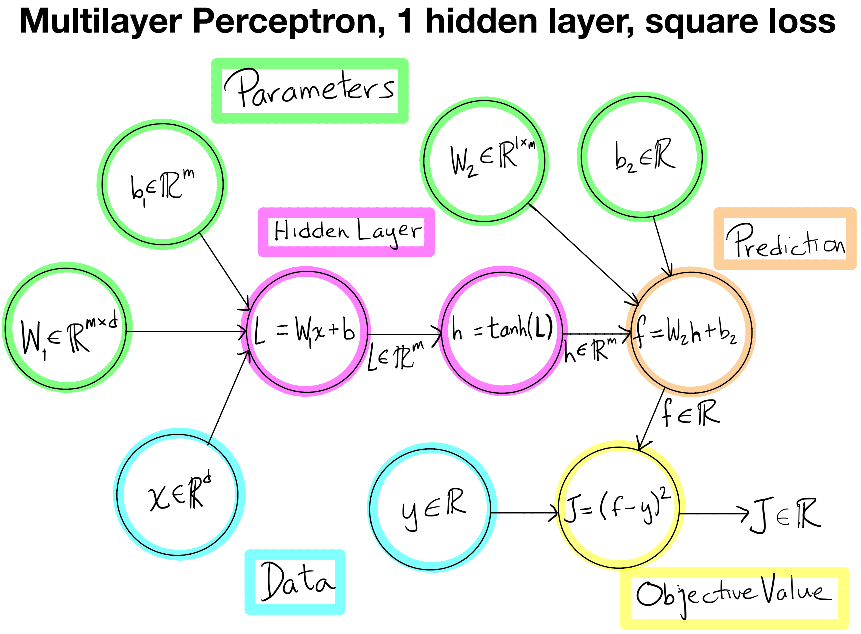
1. Complete the class `L2NormPenaltyNode` in `nodes.py`. If your code is correct, you should be able to pass test_L2NormPenaltyNode in `ridge_regression.t.py`. Please attach a screenshot that shows the test results for this question.

2. Complete the class `SumNode` in `nodes.py`. If your code is correct, you should be able to pass test_SumNode in `ridge_regression.t.py`. Please attach a screenshot that shows the test results for this question.

3. Implement ridge regression with w regularized and b unregularized. Do this by completing the `__init__` method in `ridge_regression.py`, using the classes created above. When complete, you should be able to pass the tests in `ridge_regression.t.py`. Report the average square error on the **training** set for the parameter settings given in the `main()` function.

4 Multilayer Perceptron

Let's now turn to a multilayer perceptron (MLP) with a single hidden layer and a square loss. We'll implement the computation graph illustrated below:



The crucial new piece here is the nonlinear **hidden layer**, which is what makes the multilayer perceptron a significantly larger hypothesis space than linear prediction functions.

4.1 The standard non-linear layer

The multilayer perceptron consists of a sequence of "layers" implementing the following non-linear function

$$h(x) = \sigma\left(Wx + b\right),$$

where $x \in \mathbb{R}^d$, $W \in \mathbb{R}^{m\times d}$, and $b \in \mathbb{R}^m$, and where m is often referred to as the number of **hidden units** or **hidden nodes**. σ is some non-linear function, typically tanh or ReLU, applied element-wise to the argument of σ. Referring to the computation graph illustration above, we will implement this nonlinear layer with two nodes, one implementing the affine transform $L = W_1 x + b_1$, and the other implementing the nonlinear function $h = \tanh(L)$. In this problem, we'll work out how to implement the backward method for each of these nodes.

The Affine Transformation

In a general neural network, there may be quite a lot of computation between any given affine transformation $Wx + b$ and the final objective function value J. We will capture all of that in a function $f : \mathbb{R}^m \to \mathbb{R}$, for which $J = f(Wx + b)$. Our goal is to find the partial derivative of J with respect to each element of W, namely $\partial J/\partial W_{ij}$, as well as the partials $\partial J/\partial b_i$, for each element of b. For convenience, let $y = Wx + b$, so we can write $J = f(y)$. Suppose we have already computed the partial derivatives of J with respect to the entries of $y = (y_1, \dots, y_m)^T$, namely $\frac{\partial J}{\partial y_i}$ for $i = 1, \dots, m$. Then by the chain rule, we have

$$\frac{\partial J}{\partial W_{ij}} = \sum_{r=1}^{m} \frac{\partial J}{\partial y_r} \frac{\partial y_r}{\partial W_{ij}}.$$

4. Show that $\frac{\partial J}{\partial W_{ij}} = \frac{\partial J}{\partial y_i} x_j$, where $x = (x_1, \dots, x_d)^T$. [Hint: Although not necessary, you might find it helpful to use the notation $\delta_{ij} = \begin{cases} 1 & i = j \\ 0 & \text{else} \end{cases}$. So, for examples, $\partial_{x_j} \left(\sum_{i=1}^{n} x_i^2 \right) = 2x_i \delta_{ij} = 2x_j$.]

   To answer this question we must calculate $\frac{\partial y_r}{\partial W_{ij}}$:

   If we nudge the entry i, j of the weight matrix W we have changes corresponding as following:

$$\frac{\partial y_r}{\partial W_{ij}} \rightarrow \begin{cases} x_j & if\ r = j \\ 0 & if\ otherwise \end{cases}$$

   This is clear to see by the definition of matrix - vector multiplication. If we have Ax = b then $b_i = \langle A_i, x \rangle$ where A_i is the i^{th} row of the matrix A. If we hold all of the other indices of the W matrix constant, that is to say $W_{r,k}$ is constant where $r, j \neq i, j$, then when we take the partial derivative of y with respect to $W_{i,j}$, those constants go to 0. Therefore, when we evaluate the derivative of our dot product $\langle A_i, x \rangle$, we get a summation of 0's with one term that is non zero, x_j:

$$\frac{\partial y_r}{\partial W_{ij}} \rightarrow \langle W_i, x \rangle = \sum_{r=1}^{d} \delta_{ij} \times x_r = x_j\ where\ \delta_{ij} \rightarrow \begin{cases} x_j & if\ r = j \\ 0 & if\ otherwise \end{cases}$$

   Therefore, what we have is:

$$\frac{\partial J}{\partial W_{ij}} = \frac{\partial J}{\partial y_i} x_j$$

5. Now let's vectorize this. Let's write $\frac{\partial J}{\partial y} \in \mathbb{R}^{m \times 1}$ for the column vector whose ith entry is $\frac{\partial J}{\partial y_i}$. Let's also define the matrix $\frac{\partial J}{\partial W} \in \mathbb{R}^{m \times d}$, whose ij'th entry is $\frac{\partial J}{\partial W_{ij}}$. Generally speaking, we'll always take $\frac{\partial J}{\partial A}$ to be an array of the same size ("shape" in numpy) as A. Give a vectorized expression for $\frac{\partial J}{\partial W}$ in terms of the column vectors $\frac{\partial J}{\partial y}$ and x. [Hint: Outer product.]

   We want a matrix $\frac{\partial J}{\partial W} \in \mathbb{R}^{m \times d}$ whose ij'th takes the form $\frac{\partial J}{\partial W_{ij}}$. If we're given $\frac{\partial J}{\partial y} \in \mathbb{R}^{m \times 1}$ then what we need is a vector in $\mathbb{R}^{1 \times d}$ to take the outer product with to create our $\frac{\partial J}{\partial W}$ matrix.

We know from the last problem that

$$\frac{\partial J}{\partial W_{ij}} = \frac{\partial J}{\partial y_i} x_j$$

Its easy to see that if we take the outer product of the vector $\frac{\partial J}{\partial y} \in \mathbb{R}^{m \times 1}$and the vector $x \in \mathbb{R}^{1 \times d}$ then we'll have a matrix

$$\frac{\partial J}{\partial y} \otimes x = \frac{\partial J}{\partial y} x^T \rightarrow (\frac{\partial J}{\partial y} \otimes x)_{i,j} = \frac{\partial J}{\partial y}_i x_j \textit{ therefore } \frac{\partial J}{\partial y} \otimes x = \frac{\partial J}{\partial W}$$

6. In the usual way, define $\frac{\partial J}{\partial x} \in \mathbb{R}^d$, whose i'th entry is $\frac{\partial J}{\partial x_i}$. Show that

$$\frac{\partial J}{\partial x} = W^T \left(\frac{\partial J}{\partial y} \right)$$

[Note, if x is just data, technically we won't need this derivative. However, in a multilayer perceptron, x may actually be the output of a previous hidden layer, in which case we will need to propagate the derivative through x as well.]

Using the chain rule, we know that Then the partial of J with respect to an x_j is:

$$\frac{\partial J}{\partial x_j} = \sum_{i=1}^{m} \frac{\partial J}{\partial y_i} \frac{\partial y_i}{\partial x_j}$$

If we consider the change of x_j as it corresponds to the output of y_i:

$$\frac{\partial y_i}{\partial x_j} = W_{ij}$$

Therefore, if were summing over m, we have:

$$\frac{\partial J}{\partial x_j} = \sum_{i=1}^{m} \frac{\partial J}{\partial y_i} W_{ij} = \langle \frac{\partial J}{\partial y}, W_j \rangle = \langle W_j, \frac{\partial J}{\partial y}, \rangle = W_j^T \frac{\partial J}{\partial y}$$

Where we can see that the partial derivative of J with respect to x_j is a linear combination of the entries of W scaled by the partial derivatives of J with respect to y_i. If we compute all of the partial derivatives at once, we arrive at our desired equivalency:

$$\frac{\partial J}{\partial x} = W^T \frac{\partial J}{\partial y}$$

7. Show that $\frac{\partial J}{\partial b} = \frac{\partial J}{\partial y}$, where $\frac{\partial J}{\partial b}$ is defined in the usual way.

The vector b, is just the linear bias term, hence its derivative will always be 1.

$$\frac{\partial y_i}{\partial b} = 1$$

Using the chain rule:

$$\frac{\partial J}{\partial b} = \sum_{i=1}^{m} \frac{\partial J}{\partial y_i} \times 1 = \frac{\partial J}{\partial y}$$

Element-wise Transformers

Our nonlinear activation function nodes take an array (e.g. a vector, matrix, higher-order tensor, etc), and apply the same nonlinear transformation $\sigma : \mathbb{R} \to \mathbb{R}$ to every element of the array. Let's abuse notation a bit, as is usually done in this context, and write $\sigma(A)$ for the array that results from applying $\sigma(\cdot)$ to each element of A. If σ is differentiable at $x \in \mathbb{R}$, then we'll write $\sigma'(x)$ for the derivative of σ at x, with $\sigma'(A)$ defined analogously to $\sigma(A)$.

Suppose the objective function value J is written as $J = f(\sigma(A))$, for some function $f : S \mapsto \mathbb{R}$, where S is an array of the same dimensions as $\sigma(A)$ and A. As before, we want to find the array $\frac{\partial J}{\partial A}$ for any A. Suppose for some A we have already computed the array $\frac{\partial J}{\partial S} = \frac{\partial f(S)}{\partial S}$ for $S = \sigma(A)$. At this point, we'll want to use the chain rule to figure out $\frac{\partial J}{\partial A}$. However, because we're dealing with arrays of arbitrary shapes, it can be tricky to write down the chain rule. Appropriately, we'll use a tricky convention: We'll assume all entries of an array A are indexed by a single variable. So, for example, to sum over all entries of an array A, we'll just write $\sum_i A_i$.

8. Show that $\frac{\partial J}{\partial A} = \frac{\partial J}{\partial S} \odot \sigma'(A)$, where we're using $\odot$ to represent the **Hadamard product**. If A and B are arrays of the same shape, then their Hadamard product $A \odot B$ is an array with the same shape as A and B, and for which $(A \odot B)_i = A_i B_i$. That is, it's just the array formed by multiplying corresponding elements of A and B. Conveniently, in `numpy` if `A` and `B` are arrays of the same shape, then `A*B` is their Hadamard product.

   We know that the derivative of the non linear transformation $\sigma(\cdot)$ is $\sigma'(\cdot)$, therefore, applying the derivative to A we have $\frac{\partial \sigma}{\partial A} = \sigma'(A)$. We want to compute $\frac{\partial J}{\partial A}$, using the chain rule we know this is equal to:

   $$\frac{\partial J}{\partial A} = \frac{\partial J}{\partial S}\frac{\partial S}{\partial A}$$

   Lets observe what happens when we change a single entry, i, from our array A.

   $$\frac{\partial S}{\partial A_i} = \frac{\partial \sigma(A)}{\partial A_i} = \sigma'(A_i)$$

   Where $S = \sigma(A)$ and $f(S)$ maps S to $\mathbb{R}$ for every element of the array. Since what we have is a an arbitrary mapping of every element of S through $f(S)$, and the mapping of $\mathbb{R}$ to $\mathbb{R}$ from $\sigma'(A)$ each element of S is a multiplication of two reals, dependent on the inputs of A_i and the output of $\sigma(A_i)$:

   $$\frac{\partial J}{\partial A_i} = \frac{\partial J}{\partial S_i}\frac{\partial S}{\partial A_i} = \frac{\partial J}{\partial S_i}\frac{\partial \sigma(A)}{\partial A_i} = \frac{\partial J}{\partial S_i}\sigma'(A) \to \frac{\partial J}{\partial A} = \frac{\partial J}{\partial S} \odot \sigma'(A)$$

   where $\odot$ is the element wise Hadamard product.

4.2 MLP Implementation

9. Complete the class `AffineNode` in `nodes.py`. Be sure to propagate the gradient with respect to x as well, since when we stack these layers, x will itself be the output of another node that depends on our optimization parameters. If your code is correct, you should be able to pass test_AffineNode in `mlp_regression.t.py`. Please attach a screenshot that shows the test results for this question.

10. Complete the class `TanhNode` in `nodes.py`. As you'll recall, $\frac{d}{dx}\tanh(x) = 1-\tanh^2 x$. Note that in the forward pass, we'll already have computed tanh of the input and stored it in self.out. So make sure to use `self.out` and not recalculate it in the backward pass. If your code is correct, you should be able to pass test_TanhNode in `mlp_regression.t.py`. Please attach a screenshot that shows the test results for this question.

11. Implement an MLP by completing the skeleton code in `mlp_regression.py` and making use of the nodes above. Your code should pass the tests provided in `mlp_regression.t.py`. Note that to break the symmetry of the problem, we initialize our weights to small random values, rather than all zeros, as we often do for convex optimization problems. Run the MLP for the two settings given in the `main()` function and report the average **training** error. Note that with an MLP, we can take the original scalar as input, in the hopes that it will learn nonlinear features on its own, using the hidden layers. In practice, it is quite challenging to get such a neural network to fit as well as one where we provide features.

4.3 Multiclass classification with an MLP (Optional)

We consider a generic classification problem with K classes over inputsn x of dimension d. Using a MLP we will compute a K-dimensional vector z representing scores,

$$z = W_2 \tanh(W_1 x + b_1) + b_2,$$

with $W_1 \in \mathbb{R}^{m\times d}$, $b_1 \in \mathbb{R}^m$, $W_2 \in \mathbb{R}^{K\times m}$ and $b_1 \in \mathbb{R}^K$. Our model assumes that x belongs to class k with probability

$$e^{z_k} / \sum_{k=1}^{K} e^{z_k},$$

which corresponds to applying a Softmax to the scores. Given this probabilistic model we can train the model by minimizing the negative log-likelihood.

12. Implement a Softmax node. We provided skeleton code for class SoftmaxNode in `nodes.py`. If your code is correct, you should be able to pass test_SoftmaxNode in `multiclass.t.py`. Please attach a screenshot that shows the test results for this question.

13. Implement a negative log-likelihood loss node for multiclass classification. We provided skeleton code for class NLLNode in `nodes.py`. The test code for this question is combined with the test code for the next question.

14. Implement a MLP for multiclass classification by completing the skeleton code in `multiclass.py`. Your code should pass the tests in test_multiclass provided in multiclass.t.py. Please attach a screenshot that shows the test results for this question.