Automatic Reverse-Mode Differentiation

William Cohen
Modified by xxx

Out xxx
Due xxx via Blackboard

1 Background: Automatic Differentiation

1.1 Why automatic differentiation is important

Most neural network packages (e.g., Torch or Theano) don't require a user to actually derive the gradient updates for a neural model. Instead they allow the user to define a *model* in a "little language", which supports common neural-network operations like matrix multiplication, "soft max", etc, and automatically derive the gradients. Typically, the user will define a *loss function* L in this sublanguage: the loss function takes as inputs the training data, current parameter values, and hyperparameters, and outputs a scalar value. From the definition of L, the system will compute the partial derivative of the loss function with respect to every parameter, using these gradients, it's straightforward bto implement gradient-based learning methods.

Going from the definition of L to its partial derivatives is called *automatic differentiation*. In this assignment you will start with a simple automatic differentiation system written in Python, and use it to implement a neural-network package.

1.2 Wengert lists

Automatic differentiation proceeds in two stages. First, function definitions $f(x_1, \ldots, x_k)$ in the sublanguage are converted to a format called a *Wengert list*. The second stage is to evaluate the function and its gradients using the Wengert list.

A *Wengert list* defines a function $f(x_1, \ldots, x_k)$ of some inputs $x_1, \ldots, x_k$. Syntactically, it is just a list of assignment statements, where the right-hand-side (RHS) of each statement is very simple: a call to a function g (where g is one of a set of primitive functions $G = \{g_1, \ldots, g_\ell\}$ supported by the sublanguage), where the arguments to g either inputs $x_1, \ldots, x_k$, or the left-hand-side (LHS) of a previous assignment. (The functions in G will be called *operators* here.) The output of the function is the LHS of the last item in the list. For example, a Wengert list for

$$f(x_1, x_2) \equiv (2x_1 + x_2)^2 \tag{1}$$

with the functions $G = \{\texttt{add}, \texttt{multiply}, \texttt{square}\}$ might be

$$\begin{aligned} z_1 &= \texttt{add}(x_1, x_1) \\ z_2 &= \texttt{add}(z_1, x_2) \\ f &= \texttt{square}(z_2) \end{aligned}$$

A Wengert list for

$$f(x) \equiv x^3$$

might be

$$\begin{aligned} z_1 &= \texttt{multiply}(x, x) \\ f &= \texttt{multiply}(z_1, x) \end{aligned}$$

The set of functions G defines the sublanguage. It's convenient if they have a small fixed number of arguments, and are differentiable with respect to all their arguments. But there's no reason that they have to be scalar functions! For instance, if G contains the appropriate matrix and vector operations, the loss for logistic regression (for example $\mathbf{x}$ with label y and weight matrix W) could be written as

$$f(\mathbf{x}, \mathbf{y}, W) \equiv \texttt{crossEntropy}(\texttt{softmax}(\mathbf{x} \cdot W), \mathbf{y}) + \texttt{frobeniusNorm}(W)$$

and be compiled to the Wengert list

$$\begin{aligned} \mathbf{z}_1 &= \texttt{dot}(\mathbf{x}, W) \\ \mathbf{z}_2 &= \texttt{softmax}(\mathbf{z}_1) \\ z_3 &= \texttt{crossEntropy}(\mathbf{z}_2, \mathbf{y}) \\ z_4 &= \texttt{frobeniusNorm}(W) \\ f &= \texttt{add}(z_3, z_4) \end{aligned}$$

This implements k-class logistic regression if $\mathbf{x}$ is a d-dimensional row vector, `dot` is matrix product, W is a $d \times k$ weight matrix, and

$$\begin{aligned}
\texttt{softmax}(\langle a_1, \ldots, a_d \rangle) &\equiv \langle \frac{e^{a_1}}{\sum_{i=1}^{d} e_i^a}, \ldots, \frac{e^{a_d}}{\sum_{i=1}^{d} e_i^a} \rangle \\
\texttt{crossEntropy}(\langle a_1, \ldots, a_d \rangle, \langle b_1, \ldots, b_d \rangle) &\equiv -\sum_{i=1}^{d} a_i \log b_i \\
\texttt{frobeniusNorm}(A) &\equiv \sqrt{\sum_{i=1}^{d} \sum_{j=1}^{k} a_{i,j}^2}
\end{aligned}$$

1.3 Backpropagation through Wengert list

We'll first discuss how to use a Wengert list, and then below, discuss how to construct one.

Given a Wengert list for f, it's obvious how to evaluate f: just step through the assignment statements in order, computing each value as needed. To be perfectly clear about this, the procedure is as follows. We will encode each assignment in Python as a nested tuple

$$(z, g, (y_1, \ldots, y_k))$$

where z is a string that names the LHS variable, g is a string that names the operator, and the y_i's are strings that name the arguments to g. So the list for the function of Equation 1 would be encoded in python as

```
[ ("z1", "add", ("x1","x2")),
  ("z2", "add", ("z1","x2")),
  ("f", "square", ("z2")) ]
```

We also store functions for each operator in a Python dictionary `G`:

```
G = { "add" : lambda a,b: a+b,
      "square": lambda a:a*a }
```

Before we evaluate the function, we will store the parameters in a dictionary `val`: e.g., to evalute f and $x_1 = 3$, $x_2 = 7$ we will initialize `val` to

```
val = { "x1" : 3, "x2" : 7 }
```

The pseudo-code to evaluate f is:

```
def eval(f)
    initialize val to the inputs at which f should be evaluated
    for (z, g, (y_1, ..., y_k)) in the list:
        op = G[g]
        val[z] = op(val[y_1],...,val[y_k])
    return the last entry stored in val.
```

Some Python hints: (1) to convert $(y_1, \ldots, y_k)$ to `(val[`y_1`],...,val[`y_k`])` you might use Python's `map` function. (2) If `args` is a length-2 Python list and `g` is a function that takes two arguments (like `G[''add'']` above) then `g(*args)` will call `g` with the elements of that list as the two arguments to `g`.

To differentiate, we will use a generalization of backpropagation (backprop). We'll assume that `eval` has already been run and `val` has been populated, and we will compute, in *reverse* order, a value `delta`(z_i) for each variable z_i that appears in the list. We initialize `delta` by setting `delta`$(f) = 1$, (where f is the string that names the function output).

Informally you can think of `delta`(z_i) as the "sensitivity" of f to the variable z_i, , at the point we're evaluating f (i.e., the point a that corresponds to the initial dictionary entries we stored in `val`). Here z_i can be intermediate variable or an input. If it's an input x that we're treating as a parameter, `delta` is the gradient of the cost function, evaluated at a: i.e., `delta`$(x) = \frac{df}{dx}(a)$.

To compute these sensitivities we need to "backpropagate" through the list: when we encounter the assignment $(z, g, (y_1, \ldots, y_k))$ we will use `delta`(z) and the derivatives of g with respect to its inputs to compute the sensitivities of the y's. We will store derivatives for each operator in a Python dictionary `DG`.

Note that if g has k inputs, then we need k partial derivatives, one for each input, so the entries in `DG` are *lists* of functions. For the functions used in this example, we'll need these entries in DG.

```
DG = { "add" : [ (lambda a,b: 1),  (lambda a,b: 1) ],
       "square": [ lambda a:2*a ] }
```

To figure out these functions we used some high school calculus rules: $\frac{d}{dx}(x + y) = 1$, $\frac{d}{dy}(x + y) = 1$ and $\frac{d}{dx}(x^2) = 2x$.

Finally, the pseudo-code to compute the deltas is below. Note that we don't just store values in `delta`: we accumulate them additively.

```
def backprop(f,val)
    initialize delta: delta[f] = 1
    for (z,g,(y_1,...,y_k)) in the list, in reverse order:
        for i = 1,...,k:
            op_i = DG[g][i]
            if delta[y_i] is not defined set delta[y_i] = 0
            delta[y_i] = delta[y_i] + delta[z] * op_i(val[y_1],...,val[y_k])
```

1.4 Examples

Let's look at Equation 1. In freshman calculus you'd just probably just do this:

$$\begin{aligned} f(x_1, x_2) &= (2x_1 + x_2)^2 = 4x_1^2 + 4x_1x_2 + x_2^2 \\ \frac{df}{dx_1} &= 8x_1 + 4x_2 \\ \frac{df}{dx_2} &= 4x_1 + 2x_2 \end{aligned}$$

Here we'll use the Wengert list, in reverse, and the chain rule. This list is

$$\begin{aligned} z_1 &= \texttt{add}(x_1, x_1) \\ z_2 &= \texttt{add}(z_1, x_2) \\ f &= \texttt{square}(z_2) \end{aligned}$$

Table 1 contains a detailed derivation of $\frac{df}{dx_1}$, where in each step we either plug in a definition of a variable in the list, or use the derivative of one of the operators (`square` or `add`). Table 2 contains an analogous derivation of $\frac{df}{dx_1}$. Notice that these derivations are nearly identical. In fact, they are very analogous to the computations carried out by the `backprop` algorithm: can you see how?

Finally Table 3 shows a slightly less detailed derivation of the second sample function, $f = x^3$. It is instructive to step through the `backprop`

Derivation Step	Reason
$\frac{df}{dx_1} = \frac{dz_2^2}{dz_2} \cdot \frac{dz_2}{dx_1}$	$f = z_2^2$
$\frac{df}{dx_1} = 2z_2 \cdot \frac{dz_2}{dx_1}$	$\frac{d(a^2)}{da} = 2a$
$\frac{df}{dx_1} = 2z_2 \cdot \frac{d(z_1+x_2)}{dx_1}$	$z_2 = z_1 + x_2$
$\frac{df}{dx_1} = 2z_2 \cdot \left(1 \cdot \frac{dz_1}{dx_1} + 1 \cdot \frac{dx_2}{dx_1}\right)$	$\frac{d(a+b)}{da} = \frac{d(a+b)}{db} = 1$
$\frac{df}{dx_1} = 2z_2 \cdot \left(1 \cdot \frac{d(x_1+x_1)}{dx_1} + 1 \cdot \frac{dx_2}{dx_1}\right)$	$z_1 = x_1 + x_1$
$\frac{df}{dx_1} = 2z_2 \cdot \left(1 \cdot \left(1 \cdot \frac{dx_1}{dx_1} + 1 \cdot \frac{dx_1}{dx_1}\right) + 1 \cdot \frac{dx_2}{dx_1}\right)$	$\frac{d(a+b)}{da} = \frac{d(a+b)}{db} = 1$
$\frac{df}{dx_1} = 2z_2 \cdot (1 \cdot (1 \cdot 1 + 1 \cdot 1) + 1 \cdot 0)$	$\frac{da}{da} = 1$ and $\frac{da}{db} = 0$ for inputs a, b
$\frac{df}{dx_1} = 2z_2 \cdot 2 = 8x_1 + 4x_2$	simplify

Table 1: A detailed derivation of $\frac{df}{dx_1}$ for $f = z_2^2; z_2 = z_1 + x_2; z_1 = x_1 + x_1$

Derivation Step	Reason
$\frac{df}{dx_2} = \frac{dz_2^2}{dz_2} \cdot \frac{dz_2}{dx_2}$	$f = z_2^2$
$\frac{df}{dx_2} = 2z_2 \cdot \frac{dz_2}{dx_2}$	$\frac{d(a^2)}{da} = 2a$
$\frac{df}{dx_2} = 2z_2 \cdot \frac{d(z_1+x_2)}{dx_2}$	$z_2 = z_1 + x_2$
$\frac{df}{dx_2} = 2z_2 \cdot \left(1 \cdot \frac{dz_1}{dx_2} + 1 \cdot \frac{dx_2}{dx_2}\right)$	$\frac{d(a+b)}{da} = \frac{d(a+b)}{db} = 1$
$\frac{df}{dx_2} = 2z_2 \cdot \left(1 \cdot \frac{d(x_1+x_1)}{dx_2} + 1 \cdot \frac{dx_2}{dx_2}\right)$	$z_1 = x_1 + x_1$
$\frac{df}{dx_2} = 2z_2 \cdot \left(1 \cdot \left(1 \cdot \frac{dx_2}{dx_2} + 1 \cdot \frac{dx_1}{dx_2}\right) + 1 \cdot \frac{dx_2}{dx_2}\right)$	$\frac{d(a+b)}{da} = \frac{d(a+b)}{db} = 1$
$\frac{df}{dx_1} = 2z_2 \cdot (1 \cdot (1 \cdot 1 + 1 \cdot 0) + 1 \cdot 1)$	$\frac{da}{da} = 1$ and $\frac{da}{db} = 0$ for inputs a, b
$\frac{df}{dx_1} = 2z_2 \cdot 2 = 4x_1 + 2x_2$	simplify

Table 2: A detailed derivation of $\frac{df}{dx_2}$ for $f = z_2^2; z_2 = z_1 + x_2; z_1 = x_1 + x_1$

Derivation Step	Reason
$\frac{df}{dx} = \frac{dz_1}{dx}x + z_1\frac{dx}{dx}$	$f = z_1 x$ and $\frac{d(ab)}{dx} = \frac{da}{dx}b + a\frac{db}{dx}$
$\frac{df}{dx_1} = \left(\frac{dx}{dx}x + x\frac{dx}{dx}\right) \cdot x + z_1\frac{dx}{dx}$	$z_1 = x \cdot x$ and $\frac{d(ab)}{dx} = \frac{da}{dx}b + a\frac{db}{dx}$
$\frac{df}{dx_1} = (x + x) \cdot x + x^2 = 3x^2$	$z_1 = x \cdot x$ and simplify

Table 3: A derivation of $\frac{df}{dx}$ for $f = z_1; z_1 = xx$

algorithm for these functions as well: for example the list $z_1 = x \cdot x; f = z_1 \cdot x$ leads to the `delta` updates.

$$
\begin{array}{rll}
\texttt{delta}[f] & = 1 & \\
\texttt{delta}[z_1] & \texttt{+=}\ \texttt{delta}[f] \cdot x = x & \text{arg 1 of } f = \texttt{mul(...)} \\
\texttt{delta}[x] & \texttt{+=}\ \texttt{delta}[f] \cdot z_1 = x^2 & \text{arg 2 of } f = \texttt{mul(...)} \\
\texttt{delta}[x] & \texttt{+=}\ \texttt{delta}[z_1] \cdot x = x^2 & \text{arg 1 of } z_1 = \texttt{mul(...)} \\
\texttt{delta}[x] & \texttt{+=}\ \texttt{delta}[z_1] \cdot x = x^2 & \text{arg 1 of } z_1 = \texttt{mul(...)}
\end{array}
$$

1.5 Discussion

What's going on here? Let's simplify for a minute and assume that the list is of the form

$$
\begin{aligned}
z_1 &= f_1(z_0) \\
z_2 &= f_2(z_1) \\
&\cdots \\
z_m &= f_m(z_{m-1})
\end{aligned}
$$

so $f = z_m = f_m(f_{m-1}(\ldots f_1(z_0)\ldots))$. We'll assume we can compute the f_i functions and their derivations f_i'. We know that one way to find $\frac{dz_m}{dz_0}$ would be to repeatedly use the chain rule:

$$
\begin{aligned}
\frac{dz_m}{dz_0} &= \frac{dz_m}{dz_{m-1}}\frac{dz_{m-1}}{dz_0} \\
&= \frac{dz_m}{dz_{m-1}}\frac{dz_{m-1}}{dz_{m-2}}\frac{dz_{m-2}}{dz_0} \\
&\cdots \\
&= \frac{dz_m}{dz_{m-1}}\frac{dz_{m-1}}{dz_{m-2}}\cdots\frac{dz_1}{dz_0}
\end{aligned}
$$

Let's take some time to unpack what this means. When we do derivations by hand, we are working symbolically: we are constructing a *symbolic representation* of the derivative function. This is an interesting problem—it's called *symbolic differentiation*—but it's not the same task as automatic differentiation. In automatic differentiation, we want instead an *algorithm* for *evaluating* the derivative function.

To simplify notation, let $h_{i,j}$ be the function $\frac{dz_i}{dz_j}$. (I'm doing this so that I can use $h_{i,j}(a)$ to denote the result of evaluating the function $\frac{dz_i}{dz_j}$ at a: $\frac{dz_i}{dz_j}(a)$ is hard to read.) Notice that there are m^2 of these $h_{i,j}$ functions—quite a few!—but for machine learing applications we won't care about most of them: typically we just care about the partial derivative of the cost function (the final variable in the list) with respect to the parameters, so we only need $h_{m,i}$ for certain i's.

Let's look at evaluating $h_{m,0}$ as some point $z_0 = a$ (say $z_0 = 53.4$). Again to simplify, define

$$\begin{aligned} a_1 &= f_1(a) \\ a_2 &= f_2(f_1(a)) \\ &\dots \\ a_m &= f_m(f_{m-1}(f_{m-2}(\dots f_1(a)\dots))) \end{aligned}$$

When we write

$$\frac{dz_m}{dz_0} = \frac{dz_m}{dz_{m-1}}\frac{dz_{m-1}}{dz_0}$$

We mean that: for all a

$$h_{m,0}(a) = f'_m(a_m) \cdot h_{m-1,1}(a)$$

That's a useful step because we have assumed we have available a routine to evaluate $f'_m(a_m)$: in the code this would be the function `DG[`f_m`][1]`. Continuing, when we write

$$\frac{dz_m}{dz_0} = \frac{dz_m}{dz_{m-1}}\frac{dz_{m-1}}{dz_{m-2}}\cdots\frac{dz_1}{dz_0}$$

it means that

$$h_{m,0}(a) = f'_m(a_m) \cdot f'_{m-1}(a_{m-1}) \cdots f'_2(a_1) \cdot f'_1(a)$$

When we execute the `backprop` code above, this is what we do: in particular we group the operations as

$$h_{m,0}(a) = \big(\big(\big((f'_m(a_m) \cdot f'_{m-1}(a_{m-1})\big) \cdot f'_{m-2}(a_{m-2})\big) \cdots f'_2(a_1)\big)\big) \cdot f'_1(a)$$

and the `delta`'s are partial products: specifically

$$\texttt{delta}[z_i] = f'_m(a_m) \cdots f'_i(a_i)$$

todo: discuss accumulation and distribution

1.6 Constructing Wengert lists

Wengert lists are useful but tedious to program in. Usually they are constructed using some sort programming language extension. You will be provided a package, `xman.py`, to construct Wengert lists from Python expressions: `xman` is short for "expression manager". Here is an example of using `xman`:

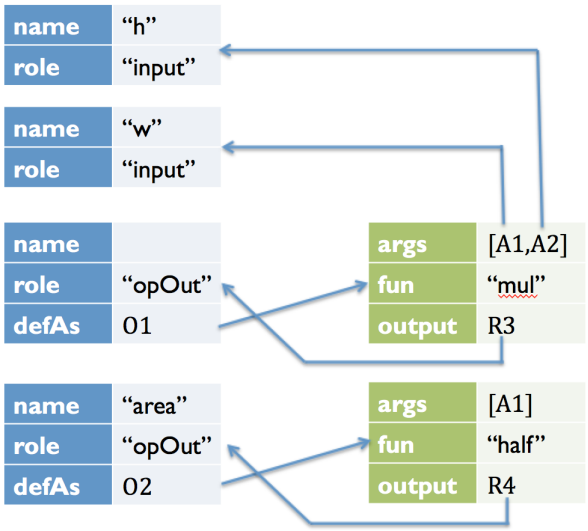
```
from xman import *
...
class f(XManFunctions):
    @staticmethod
    def half(a):
        ...
class Triangle(XMan):
    h = f.input()
    w = f.input()
    area = f.half(h*w)
...
xm = Triangle().setup()
print xm.operationSequence(xm.area)
```

In the definition of `Triangle`, the variables `h`, `w`, and `area` are called *registers*. Note that after creating an instance of a subclass of `xman.XMan`, you need to call `setup()`, which returns the newly-created instance. After the `setup` you can call the `operationSequence` method to construct a Wengert list, which will be encoded in Python as

```
[('z1', 'mul', ['h', 'w']),
 ('area', 'half', ['z1'])]
```

Internally this works as follows. There are two types of Python objects, called *Registers* and `Operations`. A `Register` corresponds to a variable, and an `Operation` corresponds to a function call.

The base `XManFunctions` class defines a method `input()` which creates a register object that is marked as an "input", meaning that it has no definition. (A similar method `param()` creates a register object that is marked as a "parameter", which like an input has no definition.) It also defines a few functions that correspond to operators, like `mul` and `add`. These return a




```
class XManFunctions(object):
    @staticmethod
    def input(default=None):
        return Register(role='input',default=default)
    ...
    @staticmethod
    def mul(a,b):
        return XManFunctions.registerDefinedByOperator('mul',a,b)
    ...
    @staticmethod
    def registerDefinedByOperator(fun,*args):
        reg = Register(role='operationOutput')
        op = Operation(fun,*args)
        reg.definedAs = op
        op.outputReg = reg
        return reg
```

Figure 1: The Python data structures created by the `Triangle` class. Blue objects are `Registers`, and green ones are `Operations`. Arrows are pointers.

register object that is cross-linked to an `Operation` object, as illustrated in Figure 1.6. (The `Register` class also uses Python's operator overloading to so that syntax like `h*w` is expanded to `XManFunctions.mul(h,w)`.)

To produce the Wengert list, the `setup()` command uses Python introspection methods to add names to each register, based on the Python variable that points to it, and generates new variable names for any reachable registers that cannot be named with Python variables. Finally, the `operationSequence` does a pre-order traversal of the data structure to create a Wengert list.

2 Assignment

*todo: more hints: return delta[z]*d, to you can do the matrix-vector thing; sometimes output is used in derivatives.*

In this assignment you will use the automatic differentiation system described above to implement two neural network architectures for character level entity classification. You will use `python` and `numpy` for this assigment. The architectures you will implement are:

- A three-layer feedforward network or Multilayer Perceptron (MLP)
- A Long Short Term Memory (LSTM) network followed by a feedforward network

Character level entity classification refers to determining the type of an entity given the characters which appear in its name as features. For example, given the name "`Noahshire`" you might guess that it is a `Location`, and given the name "`Grebulons`"[1] you might guess that it is a `Species`. Both of these are not real entities but highlight the possibility of using characters to classify entity types.

[1] `https://en.wikipedia.org/wiki/List_of_races_and_species_in_The_Hitchhiker\%27s_Guide_to_the_Galaxy#Grebulons`

3 Neural Networks

3.1 Mutlilayer Perceptron

MLPs are the simplest neural network architecture consisting of multiple layers, each of which apply a linear transformation followed by non-linear mapping to their inputs:

$$\begin{aligned} o_i &= f(\sum_{j=1}^{d_{in}} w_{ij}x_j + b_i) \\ &= f(\mathbf{x}^T\mathbf{w_i} + b_i) \end{aligned}$$

Here $\mathbf{x} \in \mathbb{R}^{d_{in}}$ is the layer input and $o_i \in \mathbb{R}$ is the i-th output of the layer. $\mathbf{w_i}$, b_i are layer parameters which we will optimize during learning. The number of outputs at each layer is called the *dimension* of that layer and we denote it by d_{out}. We would like to use vector/matrix multiplications wherever possible to utilize their fast implementation in `numpy`, and combine the above for all i as:

$$\mathbf{o} = f(\mathbf{x}^T W + \mathbf{b})^T$$

$W = [w_1, w_2, \ldots, w_{d_{out}}] \in \mathbb{R}^{d_{in} \times d_{out}}$ stacks all the weight vectors horizontally, and $\mathbf{b} \in \mathbb{R}^{d_{out}}$ holds all the biases. The non-linearity f is applied elementwise.

To further speed-up the computation we can process a minibatch of inputs together. Let $X \in \mathbb{R}^{d_{in} \times N}$ be a matrix holding N examples row-wise. We can compute the layer outputs for all of these together:

$$O = f(X^T W + B)^T \tag{2}$$

$B = \mathbf{1} \otimes \mathbf{b}^T \in \mathbb{R}^{N \times d_{out}}$ is a "broadcasted" version of the bias of appropriate dimensions. For this assigment you will use `numpy` for all matrix operations, which takes care of broadcasting automatically (see here[2] for details), hence we can use the vector $\mathbf{b}$ directly:

```
import numpy as np
Y = np.dot(X.transpose(), W) + b
O = elemwise_nonlinear_func(Y).transpose()
```

[2] `https://docs.scipy.org/doc/numpy/user/basics.broadcasting.html`

In this assigment the nonlinearity we will use is the `Rectified Linear Unit (ReLU)`:

$$f(x) = \begin{cases} x & x > 0 \\ 0 & x \leq 0 \end{cases}$$

For vectors the nonlinearity is applied element-wise, and we can again use numpy broadcasting for this:

```
def relu(X):
    return np.maximum(0,X) # X can be arbitrary shape
```

We can set the size of the last layer of the MLP to produce a vector the same size as the number of labels L in our dataset. The operations described thus far map inputs to positive reals, but for classification tasks we are interested in obtaining a *distribution* over class labels. This is usually done by passing the output of MLP through a `softmax` layer:

$$p_j = \frac{e^{o_j}}{\sum_{j'=1}^{L} e^{o_{j'}}} \tag{3}$$

Note that $\mathbf{p}$ defines a valid distribution, and elements of $\mathbf{o}$ which have a high relative value will have a high probability in $\mathbf{p}$. The above equation computes $\mathbf{p}$ for a single $\mathbf{o}$, but in your code you should use `numpy` operations to compute a minibatch of distributions P from a minibatch of outputs O.

Lastly, we need to define a loss function which measures how far the distributions P output from our network are from the target distributions T. We will use the cross-entropy loss for this:

$$J = -\frac{1}{N} \sum_{i=1}^{N} \sum_{j=1}^{L} T_{ij} \log P_{ij} \tag{4}$$

Now we can take gradients of J wrt to the parameters of the network and perform Stochastic Gradient Descent (SGD).

To summarize, the first architecture you will implement for this assignment consists of an MLP with one hidden layer, followed by a softmax layer 3 and cross-entropy loss 4. Given an input X and their associated targets T,

the output is computed as:

$$
\begin{aligned}
O^1 &= \text{relu}(X^T W^1 + \mathbf{b}^1) \\
Y &= \text{relu}(O^{1T} W^2 + \mathbf{b}^2) \\
P &= \text{softmax}(Y) \\
J &= \text{cross-entropy}(T, P)
\end{aligned}
$$

3.2 Long Short Term Memory

The MLP is a powerful model – with enough hidden units it can approximate any output, but it is not the most appropriate model when the input is a sequence. For sequences, the input size of the MLP and consequently size of W^1, will increase linearly with the size of the length of the sequence and might get prohibitively large. Instead, we would like to have a model which can loop over the input sequence, and starting from an initial state iteratively updates this state based on the input at that time step. LSTMs are one example of such a model [3].

Lets say we have a sequence of inputs $x_1, x_2, \ldots, x_M \in \mathbb{R}^{d_{in}}$, an initial *cell state* $c_0 \in \mathbb{R}^{d_{out}}$ and an initial *output* $h_0 \in \mathbb{R}^{d_{out}}$. At time t the LSTM does the following updates:

$$
\begin{aligned}
i_t &= \sigma(x_t^T W_i + h_{t-1}^T U_i + b_i) \\
f_t &= \sigma(x_t^T W_f + h_{t-1}^T U_f + b_f) \\
o_t &= \sigma(x_t^T W_o + h_{t-1}^T U_o + b_o) \\
\tilde{c}_t &= tanh(x_t^T W_c + h_{t-1}^T U_c + b_c) \\
c_t &= f_t \odot c_{t-1} + i_t \odot \tilde{c}_t \\
h_t &= o_t \odot tanh(c_t)
\end{aligned}
$$

Here $W_* \in \mathbb{R}^{d_{in} \times d_{out}}$, $U_* \in \mathbb{R}^{d_{out} \times d_{out}}$, $b_* \in \mathbb{R}^{d_{out}}$ are parameters, $\odot$ is an element-wise product and σ is the sigmoid function:

$$
\sigma(x) = \frac{1}{1 + e^{-x}} \tag{5}
$$

Suppose we have a function `_step` which takes x_t, h_{t-1}, c_{t-1} and all the parameters W_*, U_*, B_* as input and returns h_t, c_t then the LSTM can be implemented as:

[3] An excellent introduction to LSTMs can be found at `http://colah.github.io/posts/2015-08-Understanding-LSTMs/`

```
def LSTM(x, params):
    h = zeros(d_out)
    c = zeros(d_out)
    for t=[1,2,...M]:
        h, c = _step(x[t], h, c, params)
    return h
```

Note: The above equations are shown for vector inputs x_t for clarity. In your implementation you must use minibatches X_t of N examples at a time, the same way as we did for the MLP.

Now we are ready to implement the second architecture for this assignment. We will replace the hidden layer for MLP in the previous section with an LSTM layer. hence the computation would be as follows:

$$
\begin{aligned}
O^1 &= \text{LSTM}(X, \text{params}) \\
Y &= \text{relu}({O^1}^T W^2 + \mathbf{b}^2) \\
P &= \text{softmax}(Y) \\
J &= \text{cross-entropy}(T, P)
\end{aligned}
$$

Where `params` is an object (python dictionary, for example) containing all the LSTM parameters.

4 Data

5 Deliverables

6 Marking breakdown