TTIC 31230 Fundamentals of Deep Learning
Problem set 1 - solution -

- Due Thursday 12:00 pm, January 12
- Zip all your ipynb&pdf file with name PS1_yourfullname to: ttic.dl.win.2017@gmail.com.
- Late Submission: submitting late work will be penalized 10% per day, maximum three days delay allowed, no submission allowed after that.

This problem sets involves understanding and modifyint the education framework (EDF). Everyone should start by installing Anaconda with Python 3.5 (https://www.continuum.io/downloads). Then open a terminal and go the problem set directory. Then enter "jupyter workbook". This should open a window in your browser from which can open PS1.ipynb (for Interactive PYthon NoteBook). You can also open the source code edf.py of the framework. It is about four pages of Python.

Problem 1. Consider the following feed-forward computation.

$$\text{Input } x$$

$$\begin{aligned} y &= g(x) \\ z &= h(x) \\ u &= f(y,z) \\ \ell &= u \end{aligned}$$

This computation can also be written as $\ell = f(g(x), h(x))$.

a. As in the slides let $x\ell$ abbreviate $\partial \ell/\partial x$. Show that backpropagation correctly calculates $x\ell$.

solution: We notate our EDF implementation as follows,

$$\begin{aligned} \mathtt{x} &= \mathtt{Param}() && (1) \\ \mathtt{y} &= \mathtt{G}(\mathtt{x}) && (2) \\ \mathtt{z} &= \mathtt{H}(\mathtt{x}) && (3) \\ \mathtt{u} &= \mathtt{F}(\mathtt{y}, \mathtt{z}) && (4) \\ \ell &= \mathtt{u} && (5) \\ & && (6) \end{aligned}$$

where `x,y,z` and `u` are instances, and `Param, G, H` and `F` are classes.

Then our backpropagation takes 6 steps as

$$\begin{aligned}
\texttt{u.grad} &= \texttt{1} \qquad \texttt{\#initialization} && (7)\\
\texttt{z.grad}, \texttt{y.grad}, \texttt{x.grad} &= 0 && (8)\\
\texttt{z.grad} &\mathrel{+}= \texttt{u.grad}\frac{\partial f}{\partial z} = \frac{\partial f}{\partial z} && (9)\\
\texttt{y.grad} &\mathrel{+}= \texttt{u.grad}\frac{\partial f}{\partial y} = \frac{\partial f}{\partial y} && (10)\\
\texttt{x.grad} &\mathrel{+}= \texttt{y.grad}\frac{\partial g}{\partial x} = \frac{\partial f}{\partial y}\frac{\partial g}{\partial x} && (11)\\
\texttt{x.grad} &\mathrel{+}= \texttt{z.grad}\frac{\partial h}{\partial x} = \frac{\partial f}{\partial z}\frac{\partial h}{\partial x}. && (12)
\end{aligned}$$

Thus,

$$\texttt{x.grad} = \frac{\partial f}{\partial y}\frac{\partial g}{\partial x} + \frac{\partial f}{\partial z}\frac{\partial h}{\partial x} = \frac{\partial \ell}{\partial x}.$$

b. Show that naive backpropagation fails for second derivatives. More specifically let $x\ell'$ abbreviate the second partial derivative $\partial^2\ell/\partial x^2$. Show that backpropagation fails for second derivatives in the sense that

$$x\ell' \neq y\ell' \frac{d^2 g}{dx^2} + z\ell' \frac{d^2 h}{dx^2}$$

solution: Naively applying our backpropagation for the second partial derivative, $\frac{\partial^2 \ell}{\partial x^2}$, we obtain the following steps.

$$\begin{aligned}
\texttt{u.grad}' &= \texttt{1} \qquad \texttt{\#initialization} && (13)\\
\texttt{z.grad}', \texttt{y.grad}', \texttt{x.grad}' &= 0 && (14)\\
\texttt{z.grad}' &\mathrel{+}= \texttt{u.grad}'\frac{\partial^2 f}{\partial z^2} && (15)\\
\texttt{y.grad}' &\mathrel{+}= \texttt{u.grad}'\frac{\partial^2 f}{\partial y^2} && (16)\\
\texttt{x.grad}' &\mathrel{+}= \texttt{y.grad}'\frac{\partial^2 g}{\partial x^2} && (17)\\
\texttt{x.grad}' &\mathrel{+}= \texttt{z.grad}'\frac{\partial^2 h}{\partial x^2} && (18)
\end{aligned}$$

where $\texttt{.grad}'$ should be the second partial derivative, $\frac{\partial^2 \ell}{\partial \cdot^2}$. Thus,

$$\begin{aligned}
\texttt{x.grad}' &= \texttt{y.grad}'\frac{\partial^2 g}{\partial x^2} + \texttt{z.grad}'\frac{\partial h^2}{\partial x^2} && (19)\\
&= \texttt{u.grad}'\frac{\partial^2 f}{\partial y^2}\frac{\partial^2 g}{\partial x^2} + \texttt{u.grad}'\frac{\partial^2 f}{\partial z^2}\frac{\partial^2 h}{\partial x^2} && (20)\\
&= \frac{\partial^2 f}{\partial y^2}\frac{\partial^2 g}{\partial x^2} + \frac{\partial^2 f}{\partial z^2}\frac{\partial^2 h}{\partial x^2} && (21)
\end{aligned}$$

On the other hand, $\frac{\partial^2 \ell}{\partial x^2}$ can be computed with simply applying the chain rule.

$$\begin{aligned}
\frac{\partial^2 \ell}{\partial x^2} &= \frac{\partial}{\partial x}\left(\frac{\partial \ell}{\partial y}\frac{\partial g}{\partial x} + \frac{\partial \ell}{\partial z}\frac{\partial h}{\partial x}\right) && (22)\\
&= \frac{\partial}{\partial x}\left(\frac{\partial \ell}{\partial y}\frac{\partial g}{\partial x}\right) + \frac{\partial}{\partial x}\left(\frac{\partial \ell}{\partial z}\frac{\partial h}{\partial x}\right). && (23)
\end{aligned}$$

For the left term,

$$\frac{\partial}{\partial x}\left(\frac{\partial \ell}{\partial y}\frac{\partial g}{\partial x}\right) = \left(\frac{\partial}{\partial x}\frac{\partial \ell}{\partial y}\right)\frac{\partial g}{\partial x} + \frac{\partial \ell}{\partial y}\frac{\partial^2 g}{\partial x^2} \qquad (24)$$

Here, $\frac{\partial \ell}{\partial y}$ is a function of y and z, then

$$\begin{aligned}
\frac{\partial}{\partial x}\frac{\partial \ell}{\partial y} &= \frac{\partial}{\partial x}\left(\left.\frac{\partial \ell(y, h(x))}{\partial y}\right|_{y=g(x)}\right) && (25)\\
&= \left(\frac{\partial}{\partial y}\frac{\partial \ell}{\partial y}\right)\frac{\partial y}{\partial x} + \left(\frac{\partial}{\partial z}\frac{\partial \ell}{\partial y}\right)\frac{\partial z}{\partial x} && (26)\\
& && (27)
\end{aligned}$$

Thus,

$$\frac{\partial}{\partial x}\left(\frac{\partial \ell}{\partial y}\frac{\partial g}{\partial x}\right) = \frac{\partial^2 \ell}{\partial y^2}\left(\frac{\partial g}{\partial x}\right)^2 + \frac{\partial^2 \ell}{\partial y \partial z}\frac{\partial h}{\partial x}\frac{\partial g}{\partial x} + \frac{\partial \ell}{\partial y}\frac{\partial^2 g}{\partial x^2}.$$

Similarly in the right term, we have

$$\begin{aligned}
\frac{\partial^2 l}{\partial x^2} &= \frac{\partial^2 \ell}{\partial y^2}\left(\frac{\partial g}{\partial x}\right)^2 + \frac{\partial \ell}{\partial y}\frac{\partial^2 g}{\partial x^2} + \frac{\partial^2 \ell}{\partial z^2}\left(\frac{\partial h}{\partial x}\right)^2 + \frac{\partial \ell}{\partial z}\frac{\partial^2 h}{\partial x^2} + 2\frac{\partial^2 \ell}{\partial y \partial z}\frac{\partial h}{\partial x}\frac{\partial g}{\partial x}\\
&\neq \texttt{x.grad}' && (28)
\end{aligned}$$

Problem 2. Give a formal derivation of the correctness of the implementation of the backward method for softmax.
solution: For a feed-forward instance

$$y = F(x)$$

the backward method should perform

$$x.\text{grad} \mathrel{\texttt{+=}} y.\text{grad}\ \nabla_x\ f(x)$$

For f a vector-to-vector function this can be written a

$$x.\text{grad} \mathrel{\texttt{+=}} y.\text{grad}\ J$$

where J is the jacobian matrix

$$J_{i,j} = \frac{\partial f_i(x)}{\partial x_j}$$

In the case of softmax we have

$$\text{Softmax}(x)_i = \frac{1}{Z}\, e^{x_i} \qquad Z = \sum_j e^{x_j}$$

Differentiating this with respect to x_j gives the following where $I[i = j]$ is the indeicator function which is 1 if $i = j$ and zero otherwise.

$$\begin{aligned}
J_{i,j} &= \left(-\frac{1}{Z^2} e^{x_i} \frac{\partial Z}{\partial x_j}\right) + \frac{1}{Z} e^{x_j} I[i = j] \\
&= (-\frac{1}{Z^2} e^{x_i} e^{x_j}) + \frac{1}{Z} e^{x_j} I[i = j] \\
&= -y.\text{value}_i \; y.value_j + y.\text{value}_j I[i = j] \\
J &= -(y.value)(y.value)^\top + D
\end{aligned}$$

Here the first term is an outer produce of v.value with itself and the second term is a diagonal matrix with the vector y.value as the diagonal values. We then have

$$\begin{aligned}
y.\text{grad}\ J &= y.\text{grad}(-(y.\text{value})(y.\text{value})^\top + D) \\
&= -(y.\text{grad} \cdot y.\text{value}) * y.\text{value} + (y.\text{grad} * y.\text{value}) \\
&= y.\text{grad} * y.\text{value} - (y.\text{grad} \cdot y.\text{value}) * y.\text{value} \\
&= y.\text{value} * (y.\text{grad} - (y.\text{grad} \cdot y.\text{value}))
\end{aligned}$$

This last expression is the one that appears in the backward method for Softmax.