Homework 1

Guilherme Albertini

September 22, 2022

Theory

Let $Linear_1 \rightarrow f \rightarrow Linear_2 \rightarrow g$ be a be a two-layer neural net architecture whereby $Linear_i(x) = \boldsymbol{W}^{(i)}\boldsymbol{x} + \boldsymbol{b}^{(i)}$ is the i^{th} affine transformation and f, g are element-wise nonlinear activation functions (else must use transposes and/or Hadamard products). When an input $\boldsymbol{x} \in \mathbb{R}^n$ is fed into the network, $\hat{\boldsymbol{y}} \in \mathbb{R}^K$ is obtained as output.

Problem 1: Regression Task

We would like to perform regression task. We choose $f(\cdot) = 5(\cdot)^+ = 5ReLU(\cdot)$ and g to be the identity function. To train, we choose MSE loss function, $\ell_{MSE}(\hat{\boldsymbol{y}}, \boldsymbol{y}) = ||(\hat{\boldsymbol{y}} - \boldsymbol{y})||^2$.

1. Name and mathematically describe the 5 programming steps you would take to train this model with PyTorch using SGD on a single batch of data.

   (a) see vid

2. For a single data point (x, y), write down all inputs and outputs for forward pass of each layer. You can only use variables and mechanics specified prior in your answer.

—$Linear_1$—
Input: $\boldsymbol{x}$
Output (z_1): $\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$
—f—
Input: $Linear_1(\boldsymbol{x}) = \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$
Output (z_2): $f(Linear_1(\boldsymbol{x})) = 5ReLU(Linear_1(\boldsymbol{x}))$
$= 5ReLU(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})$
$= 5\max(0, \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})$
—$Linear_2$—
Input: $f(Linear_1(\boldsymbol{x}))$
Output (z_3): $Linear_2(f(Linear_1(\boldsymbol{x}))) = \boldsymbol{W}^{(2)}\boldsymbol{f(Linear_1(x))} + \boldsymbol{b}^{(2)}$
$= 5\boldsymbol{W}^{(2)}((\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})) + \boldsymbol{b}^{(2)}$
—g—
Input: $Linear_2(f(Linear_1(\boldsymbol{x})))$
Output: $g(Linear_2(f(Linear_1(\boldsymbol{x})))) =$
$5(\boldsymbol{W}^{(2)}(\max(0, \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})) + \boldsymbol{b}^{(2)})\boldsymbol{I} =$
$5(\boldsymbol{W}^{(2)}(\max(0, \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})) + \boldsymbol{b}^{(2)}) = \hat{\boldsymbol{y}}$
—Loss—
Input: $\hat{\boldsymbol{y}}$
Output: $(5(\boldsymbol{W}^{(2)}(\max(0, \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})) + \boldsymbol{b}^{(2)}) - \boldsymbol{y})^2$

3. Write down the gradients calculated from the backward pass. You can only use the following variables: $\boldsymbol{x, y, W^{(1)}, b^{(1)}, W^{(2)}, b^{(2)}}$, $\frac{\partial \ell}{\partial \hat{\boldsymbol{y}}}$, $\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}$, $\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}$, where $\boldsymbol{z_1, z_2, z_3, \hat{y}}$ are outputs of $Linear_1, f, Linear_2, g$, respectively.

$$
\begin{aligned}
\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{W^{(2)}}} &= \frac{\partial(5\boldsymbol{W^{(2)}}\,\mathbf{max}(\mathbf{0}, \boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}}) + \boldsymbol{b^{(2)}})}{\partial \boldsymbol{W^{(2)}}} \\
&= 5\,\mathbf{max}(\mathbf{0}, \boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}}) \\
\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{b^{(2)}}} &= \frac{\partial(5\boldsymbol{W^{(2)}}\,\mathbf{max}(\mathbf{0}, \boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}}) + \boldsymbol{b^{(2)}})}{\partial \boldsymbol{b^{(2)}}} \\
&= 1 \\
\frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{W^{(2)}}} &= \frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{\hat{y}}}\frac{\partial \boldsymbol{\hat{y}}}{\partial \boldsymbol{z_3}}\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{W^{(2)}}} \\
&= 5\frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{\hat{y}}}\frac{\partial \boldsymbol{\hat{y}}}{\partial \boldsymbol{z_3}}\,\mathbf{max}(\mathbf{0}, \boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}}) \\
\frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{b^{(2)}}} &= \frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{\hat{y}}}\frac{\partial \boldsymbol{\hat{y}}}{\partial \boldsymbol{z_3}}\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{b^{(2)}}} \\
&= \frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{\hat{y}}}\frac{\partial \boldsymbol{\hat{y}}}{\partial \boldsymbol{z_3}}
\end{aligned}
$$

$$
\begin{aligned}
\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{z_2}} &= \frac{\partial(\boldsymbol{W^{(2)}}5\max(0, \boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}}))}{\partial 5\max(0, \boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}})} = \boldsymbol{W^{(2)}} \\
\frac{\partial \boldsymbol{z_1}}{\partial \boldsymbol{W^{(1)}}} &= \frac{\partial(\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}})}{\boldsymbol{W^{(1)}}} = x \\
\frac{\partial \boldsymbol{z_1}}{\partial \boldsymbol{b^{(1)}}} &= \frac{\partial(\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}})}{\boldsymbol{b^{(1)}}} = 1 \\
\frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{W^{(1)}}} &= \frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{\hat{y}}}\frac{\partial \boldsymbol{\hat{y}}}{\partial \boldsymbol{z_3}}\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{z_2}}\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}\frac{\partial \boldsymbol{z_1}}{\partial \boldsymbol{W^{(1)}}} \\
&= \frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{\hat{y}}}\frac{\partial \boldsymbol{\hat{y}}}{\partial \boldsymbol{z_3}}\boldsymbol{W^{(2)}}\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}\boldsymbol{x} \\
\frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{b^{(1)}}} &= \frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{\hat{y}}}\frac{\partial \boldsymbol{\hat{y}}}{\partial \boldsymbol{z_3}}\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{z_2}}\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}\frac{\partial \boldsymbol{z_1}}{\partial \boldsymbol{b^{(1)}}} \\
&= \frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{\hat{y}}}\frac{\partial \boldsymbol{\hat{y}}}{\partial \boldsymbol{z_3}}\boldsymbol{W^{(2)}}\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}
\end{aligned}
$$

4. Show the elements of $\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}, \frac{\partial \boldsymbol{\hat{y}}}{\partial \boldsymbol{z_3}}, \frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{\hat{y}}}$. Be careful about dimensionality.

Note: $i \in \{1, \ldots, m\}$ used throughout.

$$\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}} = \frac{\partial(5\max(0, \boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}}))}{\partial(\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}})}$$

$$\left(\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}}\right)_{ii} = \begin{cases} 0, & z_{1i} < 0 \\ 5, & z_{1i} > 0 \\ \text{undefined (or assigned a value 0 in code)}, & z_{1i} = 0 \end{cases}$$

Note: $\boldsymbol{z_1} = \boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}}$ and the above is a diagonal matrix

$$\left(\frac{\partial \boldsymbol{\hat{y}}}{\partial \boldsymbol{z_3}}\right)_{ii} = 1 \text{ (and 0 elsewhere, off of diagonal; an identity matrix)}$$

$$\frac{\partial \ell}{\partial \boldsymbol{\hat{y}}} = \frac{\partial(||\boldsymbol{\hat{y}} - \boldsymbol{y}||^{\boldsymbol{2}})}{\partial \boldsymbol{\hat{y}}} = 2(\boldsymbol{\hat{y}} - \boldsymbol{y})^{\boldsymbol{T}} \text{ (A vector)}$$

Problem 2: Classification Task

We would like to perform multi-class classification task, so we set $f = tanh$ and $g = \sigma$, the logistic sigmoid function, $\sigma(z) = \frac{1}{1+\exp(-z)}$.

1. If you want to train this network, what do you need to change in the equations of (1.2), (1.3) and (1.4), assuming we are using the same MSE loss function.

—f—
Input: $Linear_1(\boldsymbol{x}) = \boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}$
Output (z_2): $f(Linear_1(\boldsymbol{x})) = \tanh(Linear_1(\boldsymbol{x}))$
$= \tanh(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)})$
$= \frac{\exp(\boldsymbol{W}^{(1)}\boldsymbol{x}+\boldsymbol{b}^{(1)})-\exp(-\boldsymbol{W}^{(1)}\boldsymbol{x}-\boldsymbol{b}^{(1)})}{\exp(\boldsymbol{W}^{(1)}\boldsymbol{x}+\boldsymbol{b}^{(1)})+\exp(-\boldsymbol{W}^{(1)}\boldsymbol{x}-\boldsymbol{b}^{(1)})}$
—$Linear_2$—
Input: $f(Linear_1(\boldsymbol{x})) = \tanh(Linear_1(x))$
Output (z_3): $Linear_2(f(Linear_1(\boldsymbol{x}))) = \boldsymbol{W}^{(2)}\boldsymbol{f(Linear_1(x))} + \boldsymbol{b}^{(2)}$
$= \boldsymbol{W^{(2)}}\,\mathbf{tanh}\,(\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}}) + \boldsymbol{b^{(2)}}$
—g—
Input: $Linear_2(f(Linear_1(\boldsymbol{x})))$
Output: $g(Linear_2(f(Linear_1(\boldsymbol{x})))) =$
$\frac{1}{1+\exp(-f(Linear_1(\boldsymbol{x})))} = \frac{1}{1+\exp(-(\boldsymbol{W^{(2)}}\,\mathbf{tanh}\,(\boldsymbol{W^{(1)}x}+\boldsymbol{b^{(1)}})+\boldsymbol{b^{(2)}}))} = \hat{\boldsymbol{y}}$

$$
\begin{aligned}
\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{W^{(2)}}} &= \frac{\partial(\boldsymbol{W^{(2)}}\,\mathbf{tanh}\,(\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}}) + \boldsymbol{b^{(2)}})}{\partial \boldsymbol{W^{(2)}}} \\
&= \tanh(\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}}) \\
\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{b^{(2)}}} &= \frac{\partial(\boldsymbol{W^{(2)}}\,\mathbf{tanh}\,(\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}}) + \boldsymbol{b^{(2)}})}{\partial \boldsymbol{b^{(2)}}} \\
&= 1 \\
\frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{W^{(2)}}} &= \frac{\partial \boldsymbol{\ell}}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}} \frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{W^{(2)}}} \\
&= \frac{\partial \boldsymbol{\ell}}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}} \tanh(\boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}}) \\
\frac{\partial \boldsymbol{\ell}}{\partial \boldsymbol{b^{(2)}}} &= \frac{\partial \boldsymbol{\ell}}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}} \frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{b^{(2)}}} \\
&= \frac{\partial \boldsymbol{\ell}}{\partial \hat{\boldsymbol{y}}} \frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}}
\end{aligned}
$$

$$\frac{\partial \boldsymbol{z_3}}{\partial \boldsymbol{z_2}} = \frac{\partial(\boldsymbol{W^{(2)}}\, \mathbf{tanh}\,(\boldsymbol{W^{(1)}}\boldsymbol{x} + \boldsymbol{b^{(1)}}) + \boldsymbol{b^{(2)}})}{\partial \tanh(\boldsymbol{W^{(1)}}\boldsymbol{x} + \boldsymbol{b^{(1)}})} = \boldsymbol{W^{(2)}}$$

$$\frac{\partial \boldsymbol{z_2}}{\partial \boldsymbol{z_1}} = \frac{\partial(\tanh(\boldsymbol{W}^{(1)}\boldsymbol{x} + \boldsymbol{b}^{(1)}))}{\partial(\boldsymbol{W^{(1)}}\boldsymbol{x} + \boldsymbol{b^{(1)}})} = \frac{2}{\cosh(2(\boldsymbol{W^{(1)}}\boldsymbol{x} + \boldsymbol{b^{(1)}})) + 1}$$

$$= 1 - \tanh(\boldsymbol{W^{(1)}}\boldsymbol{x} + \boldsymbol{b^{(1)}})^2$$

$$\frac{\partial \hat{\boldsymbol{y}}}{\partial z_3} = \frac{\partial(1 + \exp(-(\boldsymbol{W^{(2)}}\, \mathbf{tanh}\,(\boldsymbol{W^{(1)}}\boldsymbol{x} + \boldsymbol{b^{(1)}}) + \boldsymbol{b^{(2)}})))^{-1}}{\partial(\tanh(W^{(1)}x + b^{(1)}) + b^{(2)})}$$

$$= \frac{\exp(-(\boldsymbol{W^{(2)}}\, \mathbf{tanh}\,(\boldsymbol{W^{(1)}}\boldsymbol{x} + \boldsymbol{b^{(1)}}) + \boldsymbol{b^{(2)}}))}{(1 + \exp(-(\boldsymbol{W^{(2)}}\, \mathbf{tanh}\,(\boldsymbol{W^{(1)}}\boldsymbol{x} + \boldsymbol{b^{(1)}}) + \boldsymbol{b^{(2)}})))^2}$$

Problem 3

Proof. □

Section 2.2

Problem 6

Blah

Problem 7

Blah

Problem 10

Blah