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) First, we initialize the parameters to random values (and can optionally save the original parameters in another variable), and tell PyTorch that we want to track their gradients.

(b) Second, we calculate the predicted values and visualise the actual and predicted values.

(c) Third,we calculate the average loss of the model using back-propagation, seeking to reduce MSE of our existing model. To do that, we compute the current gradients to approximate how to change update the existing parameters.

(d) Fourth, we update the paramaters via the chosen learning rate.

(e) Finally, iterate until early stopping or another criterion (i.e certain number of epochs, target MSE threshold, etc.) is applied. IIt is best to monitor the training and validation losses and chosen metrics when deciding to stop.

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

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

where $\boldsymbol{z_1}, \boldsymbol{z_2}, \boldsymbol{z_3}, \hat{\boldsymbol{y}}$ are outputs of $Linear_1, f, Linear_2, g$, respectively.

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

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

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

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

$$= \begin{cases} 0, & z_1 < 0 \\ 5, & z_1 > 0 \\ \text{undefined (or assigned a value in code)}, & z_1 = 0 \end{cases}$$

Note: $\boldsymbol{z_1} = \boldsymbol{W^{(1)}x} + \boldsymbol{b^{(1)}}$

$$\frac{\partial \hat{\boldsymbol{y}}}{\partial \boldsymbol{z_3}} = 1 \text{ (Due to identity transformation)}$$

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

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 (b), (c) and (d), assuming we are using the same MSE loss function.

   ya

Problem 3

Proof. □

Section 2.2

Problem 6

Blah

Problem 7

Blah

Problem 10

Blah