NNDL Exercise set 1 – Mathematical exercises

Basic definitions

1.

The leaky ReLU is defined as

$$\psi(x) = \begin{cases} x, & \text{if } x \geq 0 \\ \alpha x, & \text{if } x < 0 \end{cases} \tag{1}$$

for $0 \leq \alpha \leq 1$.

a)

Setting $\alpha = 0$ we get the basic ReLU

$$\psi(x) = \begin{cases} x, & \text{if } x \geq 0 \\ 0, & \text{if } x < 0. \end{cases} \tag{2}$$

b)

Setting $\alpha = 1$ we get

$$\begin{aligned} \psi(x) &= \begin{cases} x, & \text{if } x \geq 0 \\ x, & \text{if } x < 0 \end{cases} \\ &= x, \end{aligned} \tag{3}$$

which is the linear activation function.

c)

Noting that $x \geq \alpha x$ for $x \geq 0$, we have $\max(x, \alpha x) = x$ when $x \geq 0$. Similarly we have $x \leq \alpha x$ for $x < 0$, thus $\max(x, \alpha x) = \alpha x$ when $x < 0$. Ergo,

$$\begin{aligned} \psi(x) &= \begin{cases} x, & \text{if } x \geq 0 \\ \alpha x, & \text{if } x < 0 \end{cases} \\ &= \max(x, \alpha x). \end{aligned} \tag{4}$$

2.

We have a multi-layer neural network with linear activation function, i.e.

$$\mathbf{y}_K = \mathbf{W}_K \mathbf{W}_{K-1} \dots \mathbf{W}_2 \mathbf{W}_1 \mathbf{x} = \mathbf{M}\mathbf{x} =: g(\mathbf{x}). \tag{5}$$

a)

Assuming $\mathbf{W}_1$ is a $m \times n$ matrix, the matrix $\mathbf{W}_2$ needs to be size $p \times m$, where p need not be equal to n, for the matrix product $\mathbf{W}_2\mathbf{W}_1$ to be defined. That is, the number of columns in $\mathbf{W}_2$ must match the number of rows in $\mathbf{W}_1$.

b)

For the neural network to be injective, the matrix $\mathbf{W}_1$ must be invertible. An invertible matrix needs to be square, thus $m = n$.

c)

The network $g(\mathbf{x})$ is injective if the matrix $\mathbf{M} = \mathbf{W}_1 \dots \mathbf{W}_K$ is invertible. Since $\mathbf{M}$ is already square, because the product of square matrices is also a square matrix, the necessary and sufficient condition is $\det \mathbf{M} \neq 0$. For square matrices A and B $\det AB = \det A \det B \implies \det \mathbf{M} = \det \mathbf{W}_1 \dots \det \mathbf{W}_K$, which is non-zero if and only if $\det \mathbf{W}_i \neq 0$, $\forall i$.

Optimization

1.

Let

$$f_1(\mathbf{w}) = ||\mathbf{w}||^2 = \sum_i w_i^2 \tag{6}$$

for $\mathbf{w} \in \mathbb{R}^n$. Noting that

$$(\nabla f_1)_j = \frac{\partial}{\partial w_j} f_1 = \frac{\partial}{\partial w_j} \sum_i w_i^2 = \underbrace{\sum_i \frac{\partial}{\partial w_j} w_i^2}_{=0,\ \text{for } i \neq j} = 2w_j, \tag{7}$$

we have

$$\nabla f_1(\mathbf{w}) = 2\mathbf{w}. \tag{8}$$

2.

The Hessian of $f_1(\mathbf{w})$ is defined as

$$(H_{f_1})_{i,j} = \frac{\partial^2}{\partial w_i \partial w_j} f_1 = \frac{\partial}{\partial w_i} 2w_j = \begin{cases} 2, & i = j \\ 0, & i \neq j \end{cases} \tag{9}$$

where we used Eq. (7). Thus

$$H(f_1(\mathbf{w})) = 2\mathbb{I}. \tag{10}$$

3.

Newton's method for $f_1(\mathbf{w}) = ||\mathbf{w}||^2$ is defined as

$$\mathbf{w}_{k+1} = \mathbf{w}_k - H(f_1(\mathbf{w}_k))^{-1}\nabla f_1(\mathbf{w}_k) \tag{11}$$

for $k > 0$. From Eq. (10) we find $H(f_1(\mathbf{w}))^{-1} = \frac{1}{2}\mathbb{I}$. Substituting this and Eq. (8) to Eq. (11) we find

$$\mathbf{w}_{k+1} = \mathbf{w}_k - \frac{1}{2}\mathbb{I}\ 2\mathbf{w}_k = \mathbf{0}. \tag{12}$$

The minimum is thus found in one step, making Newton's method extremely effective. This is because $f_1(\mathbf{w})$ is a positive definite quadratic function, and thus convex.

4.

Let

$$f_0(\mathbf{w}) = \mathbf{w}^T\mathbf{z} = \sum_i w_i z_i \tag{13}$$

for $\mathbf{w}, \mathbf{z} \in \mathbb{R}^n$. Noting that

$$(\nabla f_0)_j = \frac{\partial}{\partial w_j} f_0 = \frac{\partial}{\partial w_j}\sum_i w_i z_i = \underbrace{\sum_i \frac{\partial}{\partial w_j} w_i z_i}_{=0,\ \text{for}\ i \neq j} = z_j, \tag{14}$$

we have

$$\nabla f_0(\mathbf{w}) = \mathbf{z}. \tag{15}$$

5.

Let

$$f_2(\mathbf{w}) = g(\mathbf{w}^T\mathbf{z}) = g\left(\sum_i w_i z_i\right) \tag{16}$$

for $\mathbf{w}, \mathbf{z} \in \mathbb{R}^n$, where $g : \mathbb{R} \to \mathbb{R}$ is differentiable. Using the chain rule, we have

$$\begin{aligned}(\nabla f_2)_j &= \frac{\partial}{\partial w_j} f_2 = \frac{\partial}{\partial w_j} g\left(\sum_i w_i z_i\right) = g'\left(\sum_i w_i z_i\right)\frac{\partial}{\partial w_j}\sum_i w_i z_i \\ &= g'\left(\sum_i w_i z_i\right) z_j\end{aligned} \tag{17}$$

and thus

$$\nabla f_2(\mathbf{w}) = g'(\mathbf{w}^T\mathbf{z})\mathbf{z}. \tag{18}$$

6.

Let

$$f_3(\mathbf{w}) = \mathrm{E}\{g(\mathbf{w}^T\mathbf{z})\}. \tag{19}$$

The stochastic gradient is then

$$\nabla f_3(\mathbf{w}) = \nabla \mathrm{E}\{g(\mathbf{w}^T\mathbf{z})\} = \mathrm{E}\{\nabla g(\mathbf{w}^T\mathbf{z})\} = E\{g'(\mathbf{w}^T\mathbf{z})\mathbf{z}\} \tag{20}$$

$$= \frac{1}{M}\sum_{i=1}^{M} g'(\mathbf{w}^T\mathbf{z}_i)\mathbf{z}_i \tag{21}$$

over some sample M. Above we used Eq. (18) and the linearity of expectation.

7.

a)

$$\mathbf{M}\mathbf{w} \tag{22}$$

b)

$$||\mathbf{w}||^2\mathbf{w} \tag{23}$$