Tutorial 12 : Deep Learning

1 Remainder

Matrix Trace

Let $A \in \mathbb{R}^{n \times n}$. Then the trace of A is defined as

$$Tr(A) \triangleq \sum_{i=1}^{n} A_{ii}.$$

Properties:

- Trace is a linear mapping.
- Trace is invariant under cyclic permutations - $Tr(ABC) = Tr(CAB) = Tr(BCA)$.
- $Tr(A) = \sum_{i=1}^{n} \lambda_i$.

External Definition of Gradient

Let $f(x) : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function for which

$$df = \langle g(x), dx \rangle.$$

Then $g(x)$ is the gradient of $f(x)$.

Element-wise Function

Consider a vector $x \in \mathbb{R}^n$ and a differentiable scalar function $\phi : \mathbb{R} \to \mathbb{R}$ whose derivative is denoted by $\phi'(x)$. We define the element-wise function $\Phi : \mathbb{R}^n \to \mathbb{R}^n$ as

$$\Phi(x) \triangleq \begin{bmatrix} \phi(x_1) \\ \phi(x_2) \\ \vdots \\ \phi(x_n) \end{bmatrix}.$$

The Jacobian of $\Phi(x)$ is given by

$$J_\Phi \triangleq \begin{bmatrix} \frac{\partial \phi(x_1)}{\partial x_1} & \cdots & \frac{\partial \phi(x_1)}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial \phi(x_n)}{\partial x_1} & \cdots & \frac{\partial \phi(x_n)}{\partial x_n} \end{bmatrix} = \begin{bmatrix} \phi'(x_1) & \cdots & 0 \\ \vdots & \ddots & \vdots \\ 0 & \cdots & \phi'(x_n) \end{bmatrix}.$$

The Jacobian is a diagonal matrix with the derivatives with respect to each coordinate on the diagonal.

2 Practice

Question 1

1. Consider the following neural network

$$y = c^T \Psi\big((W + V)\Phi(Wx + b)\big),$$

where $y \in \mathbb{R}$, $x, b, c \in \mathbb{R}^n$, $W, V \in \mathbb{R}^{n\times n}$ and Ψ and Φ are elementwise functions. Compute the gradient of y with respect to W.

Solution

Recall that the external definition of the gradient g is given by

$$dy = \langle g, dW\rangle.$$

Notice that W is a matrix and hence also dW which implies that dy is an inner product of matrices

$$dy = Tr(g^T dW).$$

To compute dy we draw the computational graph and we define the following auxiliary variables according to it

$$\begin{aligned}
z_0 &= Wx \\
z_1 &= z_0 + b \\
z_2 &= \Phi(z_1) \\
Z_3 &= W + V \\
z_4 &= Z_3 z_2 \\
z_5 &= \Psi(z_4) \\
y &= c^T z_5.
\end{aligned}$$

We apply the differential operator on each of the equations:

$$\begin{array}{rcc}
z_0 = Wx & \Rightarrow & dz_0 = dWx \\
z_1 = z_0 + b & \Rightarrow & dz_1 = dz_0 \\
z_2 = \Phi(z_1) & \Rightarrow & dz_2 = \Phi'(z_1)dz_1 \\
Z_3 = W + V & \Rightarrow & dZ_3 = dW \\
z_4 = Z_3 z_2 & \Rightarrow & dz_4 = dZ_3 z_2 + Z_3 dz_2 \\
z_5 = \Psi(z_4) & \Rightarrow & dz_5 = \Psi'(z_4)dz_4 \\
y = c^T z_5 & \Rightarrow & dy = c^T dz_5,
\end{array}$$

where we neglect terms such as db and dV which are constant with respect to W. Now we "back-propagate":

$$
\begin{aligned}
dy &= c^T d_z 5 \\
&= c^T \Psi'(z_4) dz_4 \\
&= c^T \Psi'(z_4)(dZ_3 z_2 + Z_3 dz_2) \\
&= c^T \Psi'(z_4) dW z_2 + c^T \Psi'(z_4) Z_3 \Phi'(z_1) dz_1 \\
&= c^T \Psi'(z_4) dW z_2 + c^T \Psi'(z_4) Z_3 \Phi'(z_1) dz_0 \\
&= c^T \Psi'(z_4) dW z_2 + c^T \Psi'(z_4) Z_3 \Phi'(z_1) dW x.
\end{aligned}
$$

Next, we use the following properties of the trace:

- $x = Tr(x) \quad \forall x \in \mathtt{R}$.
- $Tr(A + B) = Tr(A) + Tr(B)$.
- $Tr(ABC) = Tr(CAB) = Tr(BCA)$.

Therefore,

$$
\begin{aligned}
dy &= c^T \Psi'(z_4) dW z_2 + c^T \Psi'(z_4) Z_3 \Phi'(z_1) dW x \\
&= Tr\Big(c^T \Psi'(z_4) dW z_2 + c^T \Psi'(z_4) Z_3 \Phi'(z_1) dW x\Big) \\
&= Tr\Big(c^T \Psi'(z_4) dW z_2\Big) + Tr\Big(c^T \Psi'(z_4) Z_3 \Phi'(z_1) dW x\Big) \\
&= Tr\Big(z_2 c^T \Psi'(z_4) dW\Big) + Tr\Big(x c^T \Psi'(z_4) Z_3 \Phi'(z_1) dW\Big) \\
&= Tr\Bigg(\Big(\underbrace{z_2 c^T \Psi'(z_4) + x c^T \Psi'(z_4) Z_3 \Phi'(z_1)}_{g^T}\Big) dW\Bigg)
\end{aligned}
$$

Finally we got that

$$
\begin{aligned}
g &= \Psi'(z_4) c z_2^T + \Phi'(z_1) Z_3^T \Psi'(z_4) c x^T \\
&= \Psi'\Big((W+V)\Psi(Wx+b)\Big) c \Psi(Wx+b)^T + \Phi'(Wx+b)(W+V)^T \Psi'\Big((W+V)\Psi(Wx+b)\Big) c x^T.
\end{aligned}
$$