Backpropagation through an affine linear layer

In the following I will explain how we can derive the parameter update formulas for an affine linear layer in a neural network. The affine linear layers takes as input a matrix $\mathbf{X} \in \mathbb{R}^{N\times D}$ and produces the output

$$\mathbf{Y} = \mathbf{XW} + \mathbf{B}$$

where $\mathbf{W} \in \mathbb{R}^{D\times M}$ denotes the weight matrix and

$$\mathbf{B} = \begin{bmatrix} \mathbf{b} \\ \vdots \\ \mathbf{b} \end{bmatrix}$$

is the bias matrix $\mathbf{B} \in \mathbb{R}^{N\times M}$ defined via a bias vector $\mathbf{b} \in \mathbb{R}^{1\times M}$. Our goal is to derive and understand the following formulas:

$$\frac{\partial L}{\partial \mathbf{X}} = \frac{\partial L}{\partial \mathbf{Y}} \mathbf{W}^T$$
$$\frac{\partial L}{\partial \mathbf{W}} = \mathbf{X}^T \frac{\partial L}{\partial \mathbf{Y}}$$
$$\frac{\partial L}{\partial \mathbf{b}} = \mathbf{1}^T \frac{\partial L}{\partial \mathbf{Y}}$$

These formulas can be used to define a backward pass method of an affine linear layer. An implementation in Python may look as follows:

```
def backward(self, dout):
    dx = np.dot(dout, self.W.T)
    self.dW = np.dot(self.x.T, dout)
    self.db = np.sum(dout, axis=0)
    return dx
```

The equations in the second and third row contain the partial derivatives of the loss function L w.r.t the parameters of the affine layer. The first row contains the backward pass for the layers prior to the affine layer. These layers may in turn use the information to derive the partial derivatives of the loss function w.r.t. to their parameters. Let us consider a concrete example of an affine linear layer with $N = 2$, $D = 2$ and $M = 3$:

$$\mathbf{X} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \end{bmatrix}, \quad \mathbf{W} = \begin{bmatrix} w_{11} & w_{12} & w_{13} \\ w_{21} & w_{22} & w_{23} \end{bmatrix}, \quad \mathbf{b} = \begin{bmatrix} b_1 & b_2 & b_3 \end{bmatrix}$$

The forward pass of this affine linear layer would be:

$$\mathbf{Y} = \begin{bmatrix} y_{11} & y_{12} & y_{13} \\ y_{21} & y_{22} & y_{23} \end{bmatrix} = \begin{bmatrix} x_{11}w_{11} + x_{12}w_{21} + b_1 & x_{11}w_{12} + x_{12}w_{22} + b_2 & x_{11}w_{13} + x_{12}w_{23} + b_3 \\ x_{21}w_{11} + x_{22}w_{21} + b_1 & x_{21}w_{12} + x_{22}w_{22} + b_2 & x_{21}w_{13} + x_{22}w_{23} + b_3 \end{bmatrix}$$

The mathematical way to derive the update formulas is to apply the chain rule in higher dimensions to appropiate functions. To this end we view the output $\mathbf{Y}$ as the image of a function y which maps vectors from the *right* input space to vectors in the output space of the affine layer.

Derivation of $\frac{\partial L}{\partial \mathbf{X}}$

Let us start with the derivation of $\frac{\partial L}{\partial \mathbf{X}}$. To this end, we define the function $y : \mathbb{R}^4 \to \mathbb{R}^6$ is as follows:

$$x = \begin{bmatrix} x_{11} \\ x_{12} \\ x_{21} \\ x_{22} \end{bmatrix} \mapsto \begin{bmatrix} x_{11}w_{11} + x_{12}w_{21} + b_1 \\ x_{11}w_{12} + x_{12}w_{22} + b_2 \\ x_{11}w_{13} + x_{12}w_{23} + b_3 \\ x_{21}w_{11} + x_{22}w_{21} + b_1 \\ x_{21}w_{12} + x_{22}w_{22} + b_2 \\ x_{21}w_{13} + x_{22}w_{23} + b_3 \end{bmatrix} := \begin{bmatrix} y_{11}(x) \\ y_{12}(x) \\ y_{13}(x) \\ y_{21}(x) \\ y_{22}(x) \\ y_{23}(x) \end{bmatrix}$$

The values w_{ij} and b_j in the components of y are constants. Instead of using the usual coordinate indices x_i and y_i in the definition of y we use the indices as they appear in the input matrix $\mathbf{X}$ and output matrix $\mathbf{Y}$ of the affine layer. This makes it easier to see the partial derivates of the loss function L w.r.t. the components of $\mathbf{X}$.

As an intermediate step, let us assume that we knew how to calculate the loss L from the output $\mathbf{Y}$ of the affine layer. If we denote this function by $h : \mathbb{R}^6 \to \mathbb{R}$ then $L(x) = h \circ y(x)$ for any $x \in \mathbb{R}^4$ where L is viewed for a moment purely as a function of x. By the chain rule in higher dimensions[1] we could derive the derivatives of the loss L w.r.t. the components of $\mathbf{X}$ as follows:

$$J_L(x) = J_h(y(x)) \cdot J_y(x)$$

or equivalently, $J_L(x)$ equals:

$$\begin{bmatrix} \frac{\partial h}{\partial y_{11}}(y(x)) & \frac{\partial h}{\partial y_{12}}(y(x)) & \frac{\partial h}{\partial y_{13}}(y(x)) & \frac{\partial h}{\partial y_{21}}(y(x)) & \frac{\partial h}{\partial y_{22}}(y(x)) & \frac{\partial h}{\partial y_{23}}(y(x)) \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial y_{11}}{\partial x_{11}}(x) & \frac{\partial y_{11}}{\partial x_{12}}(x) & \frac{\partial y_{11}}{\partial x_{21}}(x) & \frac{\partial y_{11}}{\partial x_{22}}(x) \\ \frac{\partial y_{12}}{\partial x_{11}}(x) & \frac{\partial y_{12}}{\partial x_{12}}(x) & \frac{\partial y_{12}}{\partial x_{21}}(x) & \frac{\partial y_{12}}{\partial x_{22}}(x) \\ \frac{\partial y_{13}}{\partial x_{11}}(x) & \frac{\partial y_{13}}{\partial x_{12}}(x) & \frac{\partial y_{13}}{\partial x_{21}}(x) & \frac{\partial y_{13}}{\partial x_{22}}(x) \\ \frac{\partial y_{21}}{\partial x_{11}}(x) & \frac{\partial y_{21}}{\partial x_{12}}(x) & \frac{\partial y_{21}}{\partial x_{21}}(x) & \frac{\partial y_{21}}{\partial x_{22}}(x) \\ \frac{\partial y_{22}}{\partial x_{11}}(x) & \frac{\partial y_{22}}{\partial x_{12}}(x) & \frac{\partial y_{22}}{\partial x_{21}}(x) & \frac{\partial y_{22}}{\partial x_{22}}(x) \\ \frac{\partial y_{23}}{\partial x_{11}}(x) & \frac{\partial y_{23}}{\partial x_{12}}(x) & \frac{\partial y_{23}}{\partial x_{21}}(x) & \frac{\partial y_{23}}{\partial x_{22}}(x) \end{bmatrix}$$

The idea behind the derivation of partial derivatives of the loss function w.r.t any parameter of a neural network is to employ the technique of backpropagation which is why we can assume that the partial derivatives $\frac{\partial h}{\partial y_{ij}}$ and the corresponding values $\frac{\partial h}{\partial y_{ij}}(y(x))$ have already been calculated. These values represent the backward pass of the next layer and they are sometimes called *upstream derivatives*. Denoting these values by $\frac{\partial L}{\partial y_{ij}}$ and combining them with the last equation yields:

$$J_L(x) = \begin{bmatrix} \frac{\partial L}{\partial x_{11}} & \frac{\partial L}{\partial x_{12}} & \frac{\partial L}{\partial x_{21}} & \frac{\partial L}{\partial x_{22}} \end{bmatrix} = \begin{bmatrix} \frac{\partial L}{\partial y_{11}} & \frac{\partial L}{\partial y_{12}} & \frac{\partial L}{\partial y_{13}} & \frac{\partial L}{\partial y_{21}} & \frac{\partial L}{\partial y_{22}} & \frac{\partial L}{\partial y_{23}} \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial y_{11}}{\partial x_{11}}(x) & \frac{\partial y_{11}}{\partial x_{12}}(x) & \frac{\partial y_{11}}{\partial x_{21}}(x) & \frac{\partial y_{11}}{\partial x_{22}}(x) \\ \frac{\partial y_{12}}{\partial x_{11}}(x) & \frac{\partial y_{12}}{\partial x_{12}}(x) & \frac{\partial y_{12}}{\partial x_{21}}(x) & \frac{\partial y_{12}}{\partial x_{22}}(x) \\ \frac{\partial y_{13}}{\partial x_{11}}(x) & \frac{\partial y_{13}}{\partial x_{12}}(x) & \frac{\partial y_{13}}{\partial x_{21}}(x) & \frac{\partial y_{13}}{\partial x_{22}}(x) \\ \frac{\partial y_{21}}{\partial x_{11}}(x) & \frac{\partial y_{21}}{\partial x_{12}}(x) & \frac{\partial y_{21}}{\partial x_{21}}(x) & \frac{\partial y_{21}}{\partial x_{22}}(x) \\ \frac{\partial y_{22}}{\partial x_{11}}(x) & \frac{\partial y_{22}}{\partial x_{12}}(x) & \frac{\partial y_{22}}{\partial x_{21}}(x) & \frac{\partial y_{22}}{\partial x_{22}}(x) \\ \frac{\partial y_{23}}{\partial x_{11}}(x) & \frac{\partial y_{23}}{\partial x_{12}}(x) & \frac{\partial y_{23}}{\partial x_{21}}(x) & \frac{\partial y_{23}}{\partial x_{22}}(x) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial L}{\partial y_{11}} & \frac{\partial L}{\partial y_{12}} & \frac{\partial L}{\partial y_{13}} & \frac{\partial L}{\partial y_{21}} & \frac{\partial L}{\partial y_{22}} & \frac{\partial L}{\partial y_{23}} \end{bmatrix} \cdot \begin{bmatrix} w_{11} & w_{12} & 0 & 0 \\ w_{12} & w_{22} & 0 & 0 \\ w_{13} & w_{23} & 0 & 0 \\ 0 & 0 & w_{11} & w_{12} \\ 0 & 0 & w_{12} & w_{22} \\ 0 & 0 & w_{13} & w_{23} \end{bmatrix}$$

Rearranging the above equation yields the formula for $\frac{\partial L}{\partial \mathbf{X}}$:

$$\frac{\partial L}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial L}{\partial x_{11}} & \frac{\partial L}{\partial x_{12}} \\ \frac{\partial L}{\partial x_{21}} & \frac{\partial L}{\partial x_{22}} \end{bmatrix} = \begin{bmatrix} \frac{\partial L}{\partial y_{11}} & \frac{\partial L}{\partial y_{12}} & \frac{\partial L}{\partial y_{13}} \\ \frac{\partial L}{\partial y_{21}} & \frac{\partial L}{\partial y_{22}} & \frac{\partial L}{\partial y_{23}} \end{bmatrix} \cdot \begin{bmatrix} w_{11} & w_{12} \\ w_{12} & w_{22} \\ w_{13} & w_{23} \end{bmatrix} = \frac{\partial L}{\partial \mathbf{Y}} \cdot \mathbf{W}^T$$

A gradient of a scalar function $f : \mathbb{R}^n \to \mathbb{R}$ is by definition a vector field $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$. Thus

[1] I have included the corresponding theorem at the end of the document.

evaluating the gradient at a point $x \in \mathbb{R}^n$ yields a vector $\nabla f(x) \in \mathbb{R}^n$. In the context of deep learning $\frac{\partial L}{\partial \mathbf{X}}$ is sometimes refered to as the gradient of the loss L w.r.t. to the input $\mathbf{X}$. As the update formula shows $\frac{\partial L}{\partial \mathbf{X}}$ is not a vector, hence not a gradient, and should therefore be refered to as a rearrangment of the components of the gradient of L w.r.t. to $\mathbf{X}$.

Derivation of $\frac{\partial L}{\partial \mathbf{W}}$ and $\frac{\partial L}{\partial \mathbf{b}}$

These formulas can be derived analogously. For example, when deriving the formula for $\frac{\partial L}{\partial \mathbf{W}}$ we need to define the function $y : \mathbb{R}^6 \to \mathbb{R}^6$ as follows:

$$w = \begin{bmatrix} w_{11} \\ w_{12} \\ w_{13} \\ w_{21} \\ w_{22} \\ w_{23} \end{bmatrix} \mapsto \begin{bmatrix} x_{11}w_{11} + x_{12}w_{21} + b_1 \\ x_{11}w_{12} + x_{12}w_{22} + b_2 \\ x_{11}w_{13} + x_{12}w_{23} + b_3 \\ x_{21}w_{11} + x_{22}w_{21} + b_1 \\ x_{21}w_{12} + x_{22}w_{22} + b_2 \\ x_{21}w_{13} + x_{22}w_{23} + b_3 \end{bmatrix} := \begin{bmatrix} y_{11}(x) \\ y_{12}(x) \\ y_{13}(x) \\ y_{21}(x) \\ y_{22}(x) \\ y_{23}(x) \end{bmatrix}$$

and proceed as for $\frac{\partial L}{\partial \mathbf{X}}$. This leads to the equation:

$$\begin{aligned} &\begin{bmatrix} \frac{\partial L}{\partial y_{11}} & \frac{\partial L}{\partial y_{12}} & \frac{\partial L}{\partial y_{13}} & \frac{\partial L}{\partial y_{21}} & \frac{\partial L}{\partial y_{22}} & \frac{\partial L}{\partial y_{23}} \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial y_{11}}{\partial w_{11}}(x) & \frac{\partial y_{11}}{\partial w_{12}}(x) & \frac{\partial y_{11}}{\partial w_{13}}(x) & \frac{\partial y_{11}}{\partial w_{21}}(x) & \frac{\partial y_{11}}{\partial w_{22}}(x) & \frac{\partial y_{11}}{\partial w_{23}}(x) \\ \frac{\partial y_{12}}{\partial w_{11}}(x) & \frac{\partial y_{12}}{\partial w_{12}}(x) & \frac{\partial y_{12}}{\partial w_{13}}(x) & \frac{\partial y_{12}}{\partial w_{21}}(x) & \frac{\partial y_{12}}{\partial w_{22}}(x) & \frac{\partial y_{12}}{\partial w_{23}}(x) \\ \frac{\partial y_{13}}{\partial w_{11}}(x) & \frac{\partial y_{13}}{\partial w_{12}}(x) & \frac{\partial y_{13}}{\partial w_{13}}(x) & \frac{\partial y_{13}}{\partial w_{21}}(x) & \frac{\partial y_{13}}{\partial w_{22}}(x) & \frac{\partial y_{13}}{\partial w_{23}}(x) \\ \frac{\partial y_{21}}{\partial w_{11}}(x) & \frac{\partial y_{21}}{\partial w_{12}}(x) & \frac{\partial y_{21}}{\partial w_{13}}(x) & \frac{\partial y_{21}}{\partial w_{21}}(x) & \frac{\partial y_{21}}{\partial w_{22}}(x) & \frac{\partial y_{21}}{\partial w_{23}}(x) \\ \frac{\partial y_{22}}{\partial w_{11}}(x) & \frac{\partial y_{22}}{\partial w_{12}}(x) & \frac{\partial y_{22}}{\partial w_{13}}(x) & \frac{\partial y_{22}}{\partial w_{21}}(x) & \frac{\partial y_{22}}{\partial w_{22}}(x) & \frac{\partial y_{22}}{\partial w_{23}}(x) \\ \frac{\partial y_{23}}{\partial w_{11}}(x) & \frac{\partial y_{23}}{\partial w_{12}}(x) & \frac{\partial y_{23}}{\partial w_{13}}(x) & \frac{\partial y_{23}}{\partial w_{21}}(x) & \frac{\partial y_{23}}{\partial w_{22}}(x) & \frac{\partial y_{23}}{\partial w_{23}}(x) \end{bmatrix} \\ &= \begin{bmatrix} \frac{\partial L}{\partial y_{11}} & \frac{\partial L}{\partial y_{12}} & \frac{\partial L}{\partial y_{13}} & \frac{\partial L}{\partial y_{21}} & \frac{\partial L}{\partial y_{22}} & \frac{\partial L}{\partial y_{23}} \end{bmatrix} \cdot \begin{bmatrix} x_{11} & 0 & 0 & x_{12} & 0 & 0 \\ 0 & x_{11} & 0 & 0 & x_{12} & 0 \\ 0 & 0 & x_{11} & 0 & 0 & x_{12} \\ x_{21} & 0 & 0 & x_{22} & 0 & 0 \\ 0 & x_{21} & 0 & 0 & x_{22} & 0 \\ 0 & 0 & x_{21} & 0 & 0 & x_{22} \end{bmatrix} \end{aligned}$$

and rearranging the above equation yields the formula for $\frac{\partial L}{\partial \mathbf{W}}$:

$$\frac{\partial L}{\partial \mathbf{W}} = \begin{bmatrix} x_{11} & x_{21} \\ x_{12} & x_{22} \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial L}{\partial y_{11}} & \frac{\partial L}{\partial y_{12}} & \frac{\partial L}{\partial y_{13}} \\ \frac{\partial L}{\partial y_{21}} & \frac{\partial L}{\partial y_{22}} & \frac{\partial L}{\partial y_{23}} \end{bmatrix} = \mathbf{X}^T \cdot \frac{\partial L}{\partial \mathbf{Y}}$$

Last but not least, defining the function $y : \mathbb{R}^3 \to \mathbb{R}^6$ as

$$b = \begin{bmatrix} b_1 \\ b_2 \\ b_3 \end{bmatrix} \mapsto \begin{bmatrix} x_{11}w_{11} + x_{12}w_{21} + b_1 \\ x_{11}w_{12} + x_{12}w_{22} + b_2 \\ x_{11}w_{13} + x_{12}w_{23} + b_3 \\ x_{21}w_{11} + x_{22}w_{21} + b_1 \\ x_{21}w_{12} + x_{22}w_{22} + b_2 \\ x_{21}w_{13} + x_{22}w_{23} + b_3 \end{bmatrix} := \begin{bmatrix} y_{11}(x) \\ y_{12}(x) \\ y_{13}(x) \\ y_{21}(x) \\ y_{22}(x) \\ y_{23}(x) \end{bmatrix}$$

we can proceed as for $\frac{\partial L}{\partial \mathbf{X}}$ and $\frac{\partial L}{\partial \mathbf{W}}$ to derive the equation

$$\begin{bmatrix} \frac{\partial L}{\partial y_{11}} & \frac{\partial L}{\partial y_{12}} & \frac{\partial L}{\partial y_{13}} & \frac{\partial L}{\partial y_{21}} & \frac{\partial L}{\partial y_{22}} & \frac{\partial L}{\partial y_{23}} \end{bmatrix} \cdot \begin{bmatrix} \frac{\partial y_{11}}{\partial b_1}(x) & \frac{\partial y_{11}}{\partial b_2}(x) & \frac{\partial y_{11}}{\partial b_3}(x) \\ \frac{\partial y_{12}}{\partial b_1}(x) & \frac{\partial y_{12}}{\partial b_2}(x) & \frac{\partial y_{12}}{\partial b_3}(x) \\ \frac{\partial y_{13}}{\partial b_1}(x) & \frac{\partial y_{13}}{\partial b_2}(x) & \frac{\partial y_{13}}{\partial b_3}(x) \\ \frac{\partial y_{21}}{\partial b_1}(x) & \frac{\partial y_{21}}{\partial b_2}(x) & \frac{\partial y_{21}}{\partial b_3}(x) \\ \frac{\partial y_{22}}{\partial b_1}(x) & \frac{\partial y_{22}}{\partial b_2}(x) & \frac{\partial y_{22}}{\partial b_3}(x) \\ \frac{\partial y_{23}}{\partial b_1}(x) & \frac{\partial y_{23}}{\partial b_2}(x) & \frac{\partial y_{23}}{\partial b_3}(x) \end{bmatrix}$$

$$= \begin{bmatrix} \frac{\partial L}{\partial y_{11}} & \frac{\partial L}{\partial y_{12}} & \frac{\partial L}{\partial y_{13}} & \frac{\partial L}{\partial y_{21}} & \frac{\partial L}{\partial y_{22}} & \frac{\partial L}{\partial y_{23}} \end{bmatrix} \cdot \begin{bmatrix} 1\,0\,0 \\ 0\,1\,0 \\ 0\,0\,1 \\ 1\,0\,0 \\ 0\,1\,0 \\ 0\,0\,1 \end{bmatrix}$$

and rearranging the terms yields the formula for $\frac{\partial L}{\partial \mathbf{b}}$:

$$\frac{\partial L}{\partial \mathbf{b}} = \mathbf{1}^T \cdot \begin{bmatrix} \frac{\partial L}{\partial y_{11}} & \frac{\partial L}{\partial y_{12}} & \frac{\partial L}{\partial y_{13}} \\ \frac{\partial L}{\partial y_{21}} & \frac{\partial L}{\partial y_{22}} & \frac{\partial L}{\partial y_{23}} \end{bmatrix}$$

where $\mathbf{1} = (1, 1)^T \in \mathbb{R}^2$.

Mathematical background

As a quick reference, I have included related material to understand the mathematical details.

Definition (Gradient). *Let $f : \mathbb{R}^n \to \mathbb{R}$ be a differentiable function. The gradient of f at a point $\mathbf{x} \in \mathbb{R}^n$ is the vector given by*

$$\nabla f(\mathbf{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1}(\mathbf{x}) \\ \frac{\partial f}{\partial x_2}(\mathbf{x}) \\ \vdots \\ \frac{\partial f}{\partial x_n}(\mathbf{x}) \end{bmatrix}$$

where $\frac{\partial f}{\partial x_j}$ denotes the partial derivative of f with respect to the i-th component of $\mathbf{x}$.

Definition (Jacobian Matrix). *Let $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$ be a differentiable function. The Jacobian matrix of $\mathbf{f}$ at a point $\mathbf{x} \in \mathbb{R}^n$ is the $m \times n$ matrix given by:*

$$J_{\mathbf{f}}(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1}(\mathbf{x}) & \frac{\partial f_1}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_1}{\partial x_n}(\mathbf{x}) \\ \frac{\partial f_2}{\partial x_1}(\mathbf{x}) & \frac{\partial f_2}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_2}{\partial x_n}(\mathbf{x}) \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1}(\mathbf{x}) & \frac{\partial f_m}{\partial x_2}(\mathbf{x}) & \cdots & \frac{\partial f_m}{\partial x_n}(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} \nabla f_1(\mathbf{x})^T \\ \nabla f_2(\mathbf{x})^T \\ \vdots \\ \nabla f_m(\mathbf{x})^T \end{bmatrix}$$

where $\frac{\partial f_i}{\partial x_j}$ denotes the partial derivative of the i-th component of $\mathbf{f}$ with respect to the j-th component of $\mathbf{x}$.

Theorem (Chain Rule in Higher Dimensions). *Let* $\mathbf{u} : \mathbb{R}^n \to \mathbb{R}^m$ *and* $\mathbf{v} : \mathbb{R}^m \to \mathbb{R}^p$ *be differentiable functions. Then* $\mathbf{w} = \mathbf{v} \circ \mathbf{u}$ *is differentiable and its Jacobian matrix at a point* $\mathbf{x} \in \mathbb{R}^n$ *is given by:*

$$J_{\mathbf{w}}(\mathbf{x}) = J_{\mathbf{v}}(\mathbf{u}(\mathbf{x})) \cdot J_{\mathbf{u}}(\mathbf{x}),$$

where $J_{\mathbf{u}}(\mathbf{x})$ *is the Jacobian matrix of* $\mathbf{u}$ *at* $\mathbf{x}$*, and* $J_{\mathbf{v}}(\mathbf{u}(\mathbf{x}))$ *is the Jacobian matrix of* $\mathbf{v}$ *at* $\mathbf{u}(\mathbf{x})$*.*

Feedback

I hope this work helped you in understanding the formulas. I would be happy to receive feedback. Feel free to send an email to *leon.holtkamp@tum.de*.