Matrix & Array Derivatives

Mark van der Wilk

Department of Computing
Imperial College London

@markvanderwilk
m.vdwilk@imperial.ac.uk

October 12, 2021

Motivation: Next level curve fitting

You have now solved Linear Regression. A key design choice was which **basis functions** to use, e.g.:

$$\boldsymbol{\phi}(x)^{\top} = \begin{bmatrix} x^3 & x^2 & x & 1 \end{bmatrix} \tag{1}$$

Motivation: Next level curve fitting

You have now solved Linear Regression. A key design choice was which **basis functions** to use, e.g.:

$$\boldsymbol{\phi}(x)^{\mathsf{T}} = \begin{bmatrix} x^3 & x^2 & x & 1 \end{bmatrix} \tag{1}$$

Instead, can we learn the basis functions?

Motivation: Next level curve fitting

You have now solved Linear Regression. A key design choice was which **basis functions** to use, e.g.:

$$\boldsymbol{\phi}(x)^\intercal = \begin{bmatrix} x^3 & x^2 & x & 1 \end{bmatrix} \tag{1}$$

Instead, can we learn the basis functions?

A **neural network** parameterises functions:

$$f_\ell(\mathbf{x}) = \sigma(\mathbf{A}_\ell \mathbf{x} + \mathbf{b}_\ell) \tag{2}$$

$$f_{NN}(\mathbf{x}) = f_L(f_{L-1}(\dots f_1(\mathbf{x}) \dots)) \tag{3}$$

Motivation: Next level curve fitting

You have now solved Linear Regression. A key design choice was which **basis functions** to use, e.g.:

$$\boldsymbol{\phi}(x)^\intercal = \begin{bmatrix} x^3 & x^2 & x & 1 \end{bmatrix} \tag{1}$$

Instead, can we learn the basis functions?

A **neural network** parameterises functions:

$$f_\ell(\mathbf{x}) = \sigma(\mathbf{A}_\ell \mathbf{x} + \mathbf{b}_\ell) \tag{2}$$

$$f_{NN}(\mathbf{x}) = f_L(f_{L-1}(\dots f_1(\mathbf{x}) \dots)) \tag{3}$$

- Parameters are $\boldsymbol{\theta} = \{\mathbf{A}_\ell, \mathbf{b}_\ell\}_{\ell=1}^{L}$.

Motivation: Next level curve fitting

You have now solved Linear Regression. A key design choice was which **basis functions** to use, e.g.:

$$\boldsymbol{\phi}(x)^\intercal = \begin{bmatrix} x^3 & x^2 & x & 1 \end{bmatrix} \tag{1}$$

Instead, can we learn the basis functions?

A **neural network** parameterises functions:

$$f_\ell(\mathbf{x}) = \sigma(\mathbf{A}_\ell \mathbf{x} + \mathbf{b}_\ell) \tag{2}$$

$$f_{NN}(\mathbf{x}) = f_L(f_{L-1}(\dots f_1(\mathbf{x}) \dots)) \tag{3}$$

- Parameters are $\boldsymbol{\theta} = \{\mathbf{A}_\ell, \mathbf{b}_\ell\}_{\ell=1}^{L}$.
- How do we differentiate w.r.t. matrices?

Derivatives of matrices/arrays

How should we find derivatives like $\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbf{x}^\intercal\mathbf{A}(\theta)\mathbf{x}$ or $\frac{\mathrm{d}}{\mathrm{d}\mathbf{A}}||\mathbf{A}\mathbf{x}-\mathbf{y}||^2$?

Derivatives of matrices/arrays

How should we find derivatives like $\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbf{x}^\intercal\mathbf{A}(\theta)\mathbf{x}$ or $\frac{\mathrm{d}}{\mathrm{d}\mathbf{A}}||\mathbf{A}\mathbf{x}-\mathbf{y}||^2$?

- By the same arguments as before (can look at differences of any array, and directional derivative arguments), we can just collect the gradients of all outputs w.r.t. all inputs:

$$\frac{\partial f_{ijkl}}{\partial x_{abc}} \tag{4}$$

Derivatives of matrices/arrays

How should we find derivatives like $\frac{\mathrm{d}}{\mathrm{d}\theta}\mathbf{x}^\top\mathbf{A}(\theta)\mathbf{x}$ or $\frac{\mathrm{d}}{\mathrm{d}\mathbf{A}}||\mathbf{A}\mathbf{x} - \mathbf{y}||^2$?

- By the same arguments as before (can look at differences of any array, and directional derivative arguments), we can just collect the gradients of all outputs w.r.t. all inputs:

$$\frac{\partial f_{ijkl}}{\partial x_{abc}} \tag{4}$$

Wouldn't it be nice if there was a chain rule?

$$\frac{\mathrm{d}f}{\mathrm{d}\theta} = \frac{\mathrm{d}f}{\mathrm{d}\mathbf{A}}\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\theta}? \qquad \text{or} \qquad \frac{\mathrm{d}f}{\mathrm{d}\mathbf{A}} = \frac{\mathrm{d}f}{\mathrm{d}\mathbf{g}}\frac{\mathrm{d}\mathbf{g}}{\mathrm{d}\mathbf{A}}?$$

Chain rule

A function of a matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ is **just a multivariate function**:

$$f(\mathbf{A}) = ||\mathbf{A}\mathbf{x} - \mathbf{y}||^2$$

$$f(A_{11}, A_{21}, \ldots, A_{M1} \ldots A_{MN}) = \sum_i (\sum_{ij} A_{ij} x_j - y_i)^2$$

Chain rule

A function of a matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ is **just a multivariate function**:

$$f(\mathbf{A}) = ||\mathbf{A}\mathbf{x} - \mathbf{y}||^2$$
$$f(A_{11}, A_{21}, \ldots, A_{M1} \ldots A_{MN}) = \sum_i (\sum_{ij} A_{ij} x_j - y_i)^2$$

We just **arrange** the numbers in a different way.

Chain rule

A function of a matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ is **just a multivariate function**:

$$f(\mathbf{A}) = ||\mathbf{A}\mathbf{x} - \mathbf{y}||^2$$

$$f(A_{11}, A_{21}, \ldots, A_{M1} \ldots A_{MN}) = \sum_i (\sum_{ij} A_{ij} x_j - y_i)^2$$

We just **arrange** the numbers in a different way. So the chain rule is the same!

$$f(\mathbf{g}) = ||\mathbf{g}||^2, \qquad \mathbf{g}(\mathbf{A}) = \mathbf{A}\mathbf{x} - \mathbf{y} \tag{5}$$

$$\frac{\partial f}{\partial A_{ij}} = \sum_k \frac{\partial f}{\partial g_k} \frac{\partial g_k}{\partial A_{ij}} \tag{6}$$

Chain rule

A function of a matrix $\mathbf{A} \in \mathbb{R}^{M \times N}$ is **just a multivariate function**:

$$f(\mathbf{A}) = ||\mathbf{A}\mathbf{x} - \mathbf{y}||^2$$
$$f(A_{11}, A_{21}, \ldots, A_{M1} \ldots A_{MN}) = \sum_i (\sum_{ij} A_{ij} x_j - y_i)^2$$

We just **arrange** the numbers in a different way. So the chain rule is the same!

$$f(\mathbf{g}) = ||\mathbf{g}||^2, \qquad \mathbf{g}(\mathbf{A}) = \mathbf{A}\mathbf{x} - \mathbf{y} \tag{5}$$
$$\frac{\partial f}{\partial A_{ij}} = \sum_k \frac{\partial f}{\partial g_k} \frac{\partial g_k}{\partial A_{ij}} \tag{6}$$

$$f(\mathbf{A}) = \mathbf{x}^{\mathsf{T}} \mathbf{A} \mathbf{x} \tag{7}$$
$$\frac{\partial f}{\partial \theta} = \sum_{jk} \frac{\partial f}{\partial A_{jk}} \frac{\partial A_{jk}}{\partial \theta} \tag{8}$$

Chain rule is not straightforward

$$f(\mathbf{A}) = \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} \qquad\qquad f(\mathbf{g}) = ||\mathbf{g}||^2,\ \mathbf{g}(\mathbf{A}) = \mathbf{A}\mathbf{x} - \mathbf{y}$$

$$\frac{\partial f}{\partial \theta} = \sum_{jk} \frac{\partial f}{\partial A_{jk}} \frac{\partial A_{jk}}{\partial \theta} \qquad\qquad \frac{\partial f}{\partial A_{ij}} = \sum_{k} \frac{\partial f}{\partial g_k} \frac{\partial g_k}{\partial A_{ij}}$$

Can we find a convenient notation like earlier?

$$\frac{\mathrm{d}f}{\mathrm{d}\theta} = \frac{\mathrm{d}f}{\mathrm{d}\mathbf{A}} \frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\theta}? \qquad\qquad \frac{\mathrm{d}f}{\mathrm{d}A} = \frac{\mathrm{d}f}{\mathrm{d}\mathbf{g}} \frac{\mathrm{d}\mathbf{g}}{\mathrm{d}\mathbf{A}}? \tag{9}$$

Chain rule is not straightforward

$$f(\mathbf{A}) = \mathbf{x}^\mathsf{T}\mathbf{A}\mathbf{x} \qquad\qquad f(\mathbf{g}) = ||\mathbf{g}||^2, \, \mathbf{g}(\mathbf{A}) = \mathbf{A}\mathbf{x} - \mathbf{y}$$

$$\frac{\partial f}{\partial \theta} = \sum_{jk} \frac{\partial f}{\partial A_{jk}} \frac{\partial A_{jk}}{\partial \theta} \qquad\qquad \frac{\partial f}{\partial A_{ij}} = \sum_{k} \frac{\partial f}{\partial g_k} \frac{\partial g_k}{\partial A_{ij}}$$

Can we find a convenient notation like earlier?

$$\frac{\mathrm{d}f}{\mathrm{d}\theta} = \frac{\mathrm{d}f}{\mathrm{d}\mathbf{A}} \frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\theta}? \qquad\qquad \frac{\mathrm{d}f}{\mathrm{d}A} = \frac{\mathrm{d}f}{\mathrm{d}\mathbf{g}} \frac{\mathrm{d}\mathbf{g}}{\mathrm{d}\mathbf{A}}? \tag{9}$$

- NOT matrix multiplication, even though both $\frac{\mathrm{d}f}{\mathrm{d}\mathbf{A}}$ and $\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\theta}$ look like matrices.

Chain rule is not straightforward

$$f(\mathbf{A}) = \mathbf{x}^{\mathsf{T}}\mathbf{A}\mathbf{x} \qquad\qquad f(\mathbf{g}) = ||\mathbf{g}||^2, \, \mathbf{g}(\mathbf{A}) = \mathbf{A}\mathbf{x} - \mathbf{y}$$

$$\frac{\partial f}{\partial \theta} = \sum_{jk} \frac{\partial f}{\partial A_{jk}} \frac{\partial A_{jk}}{\partial \theta} \qquad\qquad \frac{\partial f}{\partial A_{ij}} = \sum_{k} \frac{\partial f}{\partial g_k} \frac{\partial g_k}{\partial A_{ij}}$$

Can we find a convenient notation like earlier?

$$\frac{\mathrm{d}f}{\mathrm{d}\theta} = \frac{\mathrm{d}f}{\mathrm{d}\mathbf{A}} \frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\theta}? \qquad\qquad \frac{\mathrm{d}f}{\mathrm{d}A} = \frac{\mathrm{d}f}{\mathrm{d}\mathbf{g}} \frac{\mathrm{d}\mathbf{g}}{\mathrm{d}\mathbf{A}}? \tag{9}$$

- NOT matrix multiplication, even though both $\frac{\mathrm{d}f}{\mathrm{d}\mathbf{A}}$ and $\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\theta}$ look like matrices. Check the shapes!
- Shape of $\frac{\mathrm{d}\mathbf{g}}{\mathrm{d}\mathbf{A}}$ isn't even a matrix!

Derivatives with Respect to Matrices

- Recall: A function $\boldsymbol{f} : \mathbb{R}^N \to \mathbb{R}^M$ has a gradient that is an $M \times N$-matrix with

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{x}} \in \mathbb{R}^{M \times N}, \qquad \mathrm{d}\boldsymbol{f}[m,n] = \frac{\partial f_m}{\partial x_n}$$

Gradient dimension: # target dimensions $\times$ # input dimensions

Derivatives with Respect to Matrices

- Recall: A function $f : \mathbb{R}^N \to \mathbb{R}^M$ has a gradient that is an $M \times N$-matrix with

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{x}} \in \mathbb{R}^{M \times N}, \qquad \mathrm{d}\boldsymbol{f}[m,n] = \frac{\partial f_m}{\partial x_n}$$

  Gradient dimension: # target dimensions $\times$ # input dimensions

- This generalizes to when the inputs (N) or targets (M) are **matrices**

Derivatives with Respect to Matrices

- Recall: A function $\boldsymbol{f} : \mathbb{R}^{N} \to \mathbb{R}^{M}$ has a gradient that is an $M \times N$-matrix with

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{x}} \in \mathbb{R}^{M \times N}, \qquad \mathrm{d}\boldsymbol{f}[m,n] = \frac{\partial f_m}{\partial x_n}$$

Gradient dimension: # target dimensions $\times$ # input dimensions

- This generalizes to when the inputs (N) or targets (M) are **matrices**
- Function $\boldsymbol{f} : \mathbb{R}^{M \times N} \to \mathbb{R}^{P \times Q}$, has a gradient that is a $(P \times Q) \times (M \times N)$ object (array)

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{X}} \in \mathbb{R}^{(P \times Q) \times (M \times N)}, \qquad \mathrm{d}\boldsymbol{f}[p,q,m,n] = \frac{\partial f_{pq}}{\partial X_{mn}}$$

Derivatives with Respect to Matrices

- Recall: A function $\boldsymbol{f} : \mathbb{R}^{N} \to \mathbb{R}^{M}$ has a gradient that is an $M \times N$-matrix with

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{x}} \in \mathbb{R}^{M\times N}, \qquad \mathrm{d}\boldsymbol{f}[m,n] = \frac{\partial f_m}{\partial x_n}$$

Gradient dimension: # target dimensions $\times$ # input dimensions

- This generalizes to when the inputs (N) or targets (M) are **matrices**
- Function $\boldsymbol{f} : \mathbb{R}^{M\times N} \to \mathbb{R}^{P\times Q}$, has a gradient that is a $(P \times Q) \times (M \times N)$ object (array)

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{X}} \in \mathbb{R}^{(P\times Q)\times(M\times N)}, \qquad \mathrm{d}\boldsymbol{f}[p,q,m,n] = \frac{\partial f_{pq}}{\partial X_{mn}}$$

Autodiff packages have similar consistency of shapes.

Example 1: Derivatives with Respect to Matrices

$$\boldsymbol{f} = \boldsymbol{A}\boldsymbol{x}\,, \quad \boldsymbol{f} \in \mathbb{R}^M, \boldsymbol{A} \in \mathbb{R}^{M\times N}, \boldsymbol{x} \in \mathbb{R}^N$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_M(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} A_{11}x_1 + A_{12}x_2 + & \cdots & + A_{1N}x_N \\ \vdots \quad\quad \vdots & \vdots & \vdots \\ A_{M1}x_1 + A_{M2}x_2 + & \cdots & + A_{MN}x_N \end{bmatrix}$$

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{A}} \in \mathbb{R}^{?}$$

Example 1: Derivatives with Respect to Matrices

$$\boldsymbol{f} = \boldsymbol{A}\boldsymbol{x}, \quad \boldsymbol{f} \in \mathbb{R}^M, \boldsymbol{A} \in \mathbb{R}^{M\times N}, \boldsymbol{x} \in \mathbb{R}^N$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_M(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} A_{11}x_1 + A_{12}x_2 + & \cdots & + A_{1N}x_N \\ \vdots \quad\quad \vdots & \vdots & \vdots \\ A_{M1}x_1 + A_{M2}x_2 + & \cdots & + A_{MN}x_N \end{bmatrix}$$

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{A}} \in \mathbb{R}^{\#\text{ target dim} \times \#\text{ input dim}} = M \times (M \times N)$$

$$\frac{d\boldsymbol{f}}{d\boldsymbol{A}} = \begin{bmatrix} \frac{\partial f_1}{\partial \boldsymbol{A}} \\ \vdots \\ \frac{\partial f_M}{\partial \boldsymbol{A}} \end{bmatrix}, \quad \frac{\partial f_i}{\partial \boldsymbol{A}} \in \mathbb{R}^{1\times(M\times N)}$$

Example 2: Derivatives with Respect to Matrices

$$f_i = \sum_{j=1}^{N} A_{ij} x_j, \quad i = 1, \ldots, M$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_i(\boldsymbol{x}) \\ \vdots \\ f_M(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} A_{11}x_1 + A_{12}x_2 + & \cdots & +A_{1N}x_N \\ \vdots \quad\quad \vdots & \vdots & \vdots \\ A_{i1}x_1 + A_{i2}x_2 & \cdots & +A_{iN}x_N \\ \vdots \quad\quad \vdots & \vdots & \vdots \\ A_{M1}x_1 + A_{M2}x_2 + & \cdots & +A_{MN}x_N \end{bmatrix}$$

$$\frac{\partial f_i}{\partial A_{iq}} = ? \qquad \frac{\partial f_i}{\partial A_{i,:}} = ? \qquad \frac{\partial f_i}{\partial A_{k\neq i,:}} = ? \qquad \frac{\partial f_i}{\partial \boldsymbol{A}} = ?$$

Example 2: Derivatives with Respect to Matrices

$$f_i = \sum_{j=1}^{N} A_{ij} x_j, \quad i = 1, \ldots, M$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_i(\boldsymbol{x}) \\ \vdots \\ f_M(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} A_{11}x_1 + A_{12}x_2 + & \cdots & + A_{1N}x_N \\ \vdots \quad\quad \vdots & \vdots & \vdots \\ A_{i1}x_1 + A_{i2}x_2 & \cdots & + A_{iN}x_N \\ \vdots \quad\quad \vdots & \vdots & \vdots \\ A_{M1}x_1 + A_{M2}x_2 + & \cdots & + A_{MN}x_N \end{bmatrix}$$

$$\frac{\partial f_i}{\partial A_{iq}} = \underbrace{x_q}_{\in \mathbb{R}} \quad \frac{\partial f_i}{\partial A_{i,:}} = ? \quad \frac{\partial f_i}{\partial A_{k \neq i,:}} = ? \quad \frac{\partial f_i}{\partial \boldsymbol{A}} = ?$$

Example 2: Derivatives with Respect to Matrices

$$f_i = \sum_{j=1}^{N} A_{ij}x_j, \quad i = 1, \ldots, M$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_i(\boldsymbol{x}) \\ \vdots \\ f_M(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} A_{11}x_1 + A_{12}x_2 + & \cdots & +A_{1N}x_N \\ \vdots \quad\quad \vdots & \vdots & \vdots \\ A_{i1}x_1 + A_{i2}x_2 & \cdots & +A_{iN}x_N \\ \vdots \quad\quad \vdots & \vdots & \vdots \\ A_{M1}x_1 + A_{M2}x_2 + & \cdots & +A_{MN}x_N \end{bmatrix}$$

$$\frac{\partial f_i}{\partial A_{iq}} = \underbrace{x_q}_{\in \mathbb{R}} \qquad \frac{\partial f_i}{\partial A_{i,:}} = \underbrace{\boldsymbol{x}^\top}_{\in \mathbb{R}^{1\times 1\times N}} \qquad \frac{\partial f_i}{\partial A_{k\neq i,:}} = ? \qquad \frac{\partial f_i}{\partial \boldsymbol{A}} = ?$$

Example 2: Derivatives with Respect to Matrices

$$f_i = \sum_{j=1}^{N} A_{ij} x_j, \quad i = 1, \ldots, M$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_i(\boldsymbol{x}) \\ \vdots \\ f_M(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} A_{11}x_1 + A_{12}x_2 + & \cdots & +A_{1N}x_N \\ \vdots \quad\quad \vdots & \vdots & \vdots \\ A_{i1}x_1 + A_{i2}x_2 & \cdots & +A_{iN}x_N \\ \vdots \quad\quad \vdots & \vdots & \vdots \\ A_{M1}x_1 + A_{M2}x_2 + & \cdots & +A_{MN}x_N \end{bmatrix}$$

$$\frac{\partial f_i}{\partial A_{iq}} = \underbrace{x_q}_{\in \mathbb{R}} \quad \frac{\partial f_i}{\partial A_{i,:}} = \underbrace{\boldsymbol{x}^\top}_{\in \mathbb{R}^{1 \times 1 \times N}} \quad \frac{\partial f_i}{\partial A_{k \neq i,:}} = \underbrace{\boldsymbol{0}^\top}_{\in \mathbb{R}^{1 \times 1 \times N}} \quad \frac{\partial f_i}{\partial \boldsymbol{A}} = ?$$

Example 2: Derivatives with Respect to Matrices

$$f_i = \sum_{j=1}^{N} A_{ij} x_j, \quad i = 1, \ldots, M$$

$$\begin{bmatrix} y_1 \\ \vdots \\ y_i \\ \vdots \\ y_M \end{bmatrix} = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_i(\boldsymbol{x}) \\ \vdots \\ f_M(\boldsymbol{x}) \end{bmatrix} = \begin{bmatrix} A_{11}x_1 + A_{12}x_2 + & \cdots & +A_{1N}x_N \\ \vdots \quad\quad \vdots & \vdots & \vdots \\ A_{i1}x_1 + A_{i2}x_2 & \cdots & +A_{iN}x_N \\ \vdots \quad\quad \vdots & \vdots & \vdots \\ A_{M1}x_1 + A_{M2}x_2 + & \cdots & +A_{MN}x_N \end{bmatrix}$$

$$\frac{\partial f_i}{\partial A_{iq}} = \underbrace{x_q}_{\in \mathbb{R}} \qquad \frac{\partial f_i}{\partial A_{i,:}} = \underbrace{\boldsymbol{x}^\top}_{\in \mathbb{R}^{1\times 1\times N}} \qquad \frac{\partial f_i}{\partial A_{k\neq i,:}} = \underbrace{\boldsymbol{0}^\top}_{\in \mathbb{R}^{1\times 1\times N}} \qquad \frac{\partial f_i}{\partial \boldsymbol{A}} = \underbrace{\begin{bmatrix} \boldsymbol{0}^\top \\ \vdots \\ \boldsymbol{x}^\top \\ \vdots \\ \boldsymbol{0}^\top \end{bmatrix}}_{\in \mathbb{R}^{1\times (M\times N)}}$$

Chain rule

- We now understand how gradients involving matrices are arranged in "multidimensional arrays" or "tensors".

Chain rule

- We now understand how gradients involving matrices are arranged in "multidimensional arrays" or "tensors".
- How do we perform the chain rule? Can we find a meaning for convenient notation like:

$$\frac{\mathrm{d}f}{\mathrm{d}\theta} = \frac{\mathrm{d}f}{\mathrm{d}\mathbf{A}}\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\theta}? \qquad \frac{\mathrm{d}f}{\mathrm{d}\theta} = \frac{\mathrm{d}f}{\mathrm{d}\mathbf{g}}\frac{\mathrm{d}\mathbf{g}}{\mathrm{d}\mathbf{A}}? \tag{10}$$

Chain rule

- We now understand how gradients involving matrices are arranged in "multidimensional arrays" or "tensors".
- How do we perform the chain rule? Can we find a meaning for convenient notation like:

$$\frac{\mathrm{d}f}{\mathrm{d}\theta} = \frac{\mathrm{d}f}{\mathrm{d}\mathbf{A}}\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\theta}? \qquad\qquad \frac{\mathrm{d}f}{\mathrm{d}\theta} = \frac{\mathrm{d}f}{\mathrm{d}\mathbf{g}}\frac{\mathrm{d}\mathbf{g}}{\mathrm{d}\mathbf{A}}? \tag{10}$$

- Recall: Function $\boldsymbol{f} : \mathbb{R}^{M \times N} \to \mathbb{R}^{P \times Q}$, has a gradient that is a $(P \times Q) \times (M \times N)$ object (tensor)

$$\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{X}} \in \mathbb{R}^{(P\times Q)\times(M\times N)}, \qquad \mathrm{d}\boldsymbol{f}[p,q,m,n] = \frac{\partial f_{pq}}{\partial X_{mn}}$$

Chain rule (2)

- Start from index notation chain rule (**always correct!**):

$$\frac{\partial f_{pq}}{\partial X_{mn}} = \sum_{rs} \frac{\partial f_{pq}}{\partial A_{rs}} \frac{\partial A_{rs}}{\partial X_{mn}}$$

Chain rule (2)

- Start from index notation chain rule (**always correct!**):

$$\frac{\partial f_{pq}}{\partial X_{mn}} = \sum_{rs} \frac{\partial f_{pq}}{\partial A_{rs}} \frac{\partial A_{rs}}{\partial X_{mn}}$$

- Like matrix multiplication, but with **vectorised** (vectors stacked column-by-column) matrices!

$$\frac{\mathrm{dvec}(\boldsymbol{f})}{\mathrm{dvec}(\boldsymbol{X})} = \frac{\mathrm{dvec}(\boldsymbol{f})}{\mathrm{dvec}(\boldsymbol{A})} \frac{\mathrm{dvec}(\boldsymbol{A})}{\mathrm{dvec}(\boldsymbol{X})}$$

Chain rule (2)

- Start from index notation chain rule (**always correct!**):

$$\frac{\partial f_{pq}}{\partial X_{mn}} = \sum_{rs} \frac{\partial f_{pq}}{\partial A_{rs}} \frac{\partial A_{rs}}{\partial X_{mn}}$$

- Like matrix multiplication, but with **vectorised** (vectors stacked column-by-column) matrices!

$$\frac{\mathrm{dvec}(\boldsymbol{f})}{\mathrm{dvec}(\boldsymbol{X})} = \frac{\mathrm{dvec}(\boldsymbol{f})}{\mathrm{dvec}(\boldsymbol{A})} \frac{\mathrm{dvec}(\boldsymbol{A})}{\mathrm{dvec}(\boldsymbol{X})}$$

- Keep track of grouping, sum over grouped indices:

$$\underbrace{\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\mathbf{X}}}_{(P \times Q) \times (M \times N)} = \underbrace{\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\mathbf{A}}}_{(P \times Q) \times (R \times S)} \cdot \underbrace{\frac{\mathrm{d}\mathbf{A}}{\mathrm{d}\mathbf{X}}}_{(R \times S) \times (M \times N)}$$

Summary: Matrix differentiation

We saw:

- Principle is the same for matrix and vector differentiation

Summary: Matrix differentiation

We saw:

- Principle is the same for matrix and vector differentiation
- Difference: Management of the numbers. It's about **convention**

Summary: Matrix differentiation

We saw:

- Principle is the same for matrix and vector differentiation
- Difference: Management of the numbers. It's about **convention**
- Mathematical principle is index notation, convention is defined

Summary: Matrix differentiation

We saw:

- Principle is the same for matrix and vector differentiation
- Difference: Management of the numbers. It's about **convention**
- Mathematical principle is index notation, convention is defined

You should be able to:

- Do the bookkeeping of matrix derivative shapes
- Compute derivatives of matrices
- Abstract complex derivatives into the well-defined chain rule.
- Describe the detailed index-wise summation for the chain rule.