Vector and Matrix Differentiation

6.S087 IAP 2021, Lecture **2**

MIT EECS

January 4, 2021

Table of Contents

1. Introduction and Conventions
2. Definitions
   - Vector Differentiation
   - Differentiation Layouts
   - Matrix Differentiation
3. Examples
   - Element-wise Proofs
4. Composition Rules

Introduction

We introduce matrix differentiation concepts and techniques up to a level commonly used in statistics and data science.

The main constraint for the set of functions that we will differentiate is that the sum of dimensions across the input and output must be at most two. (Why?)

Introduction

We introduce matrix differentiation concepts and techniques up to a level commonly used in statistics and data science.

The main constraint for the set of functions that we will differentiate is that the sum of dimensions across the input and output must be at most two. (Why?)

Some examples of possible functions are:

$$\begin{aligned} f : \mathbb{R} \to \mathbb{R}^n &\text{ e.g. parameter of symmetric Dirichlet} \\ f : \mathbb{R}^n \to \mathbb{R} &\text{ e.g. vector norm} \\ f : \mathbb{R}^{n \times n} \to \mathbb{R} &\text{ e.g. determinant} \\ f : \mathbb{R}^n \to \mathbb{R}^n &\text{ e.g. linear transformations} \end{aligned}$$

Conventions

We will use:

- a small letter (such as x and y) to refer to scalars
- a small letter with an arrow (such as $\vec{x}$ or $\vec{y}$) to refer to vectors
- a capital letter (such as X or Y) to refer to matrices

Conventions

We will use:

- a small letter (such as x and y) to refer to scalars
- a small letter with an arrow (such as $\vec{x}$ or $\vec{y}$) to refer to vectors
- a capital letter (such as X or Y) to refer to matrices

All vectors in this article are column vectors unless otherwise specified, meaning that entries are stacked on top of each other; they could also be seen as matrices with only one column.

Conventions

It is often more convenient to write $\frac{dx^2}{dx}$ in place of $\frac{df}{dx}$, where $f = x^2$.

In statistics in particular, we often write variables in terms of another, e.g. $\vec{y} = 2\vec{x}$ (we say that $\vec{y}$ is parametrized by $\vec{x}$); then we might consider the derivative of $\vec{y}$ with respect to $\vec{x}$, which is written $\frac{d\vec{y}}{d\vec{x}}$.

Conventions

It is often more convenient to write $\frac{dx^2}{dx}$ in place of $\frac{df}{dx}$, where $f = x^2$.

In statistics in particular, we often write variables in terms of another, e.g. $\vec{y} = 2\vec{x}$ (we say that $\vec{y}$ is parametrized by $\vec{x}$); then we might consider the derivative of $\vec{y}$ with respect to $\vec{x}$, which is written $\frac{d\vec{y}}{d\vec{x}}$.

For the rest of this lecture, we will use u, v, y to indicate variables parametrized by x, and a, b to indicate constant variables.

Vector Differentiation

The simplest form of multivariable differentiation, vector differentiation generalizes the one-dimensional concept of a derivative to functions with vector-valued inputs or outputs.

We develop the concept of the gradient by generalizing the limit definition of the (single-variable) derivative, which is

$$\lim_{t \to 0} \frac{f(x+t) - f(x)}{t} = f'(x),$$

to functions where the input is a vector.

In the multivariable case, what $t \to 0$ means is less clear, as there are many directions in which one could approach a point in $\mathbb{R}^n$.

Gradient

Given a vector $\vec{d}$ with the same dimension as $\vec{x}$, we could consider the limit

$$\nabla f(\vec{x})[\vec{d}] := \lim_{t \to 0} \frac{f(\vec{x} + t\vec{d}) - f(\vec{x})}{t},$$

which may be thought of as a function of both $\vec{x}$ and $\vec{d}$.

If we want a definition for the multidimensional derivative $\frac{df}{d\vec{x}}$ at a given point $\vec{x}$, it should not depend on d.

Gradient

It turns out, assuming that the function f is differentiable, that there exists a vector ∇f such that $\nabla f(\vec{x})[\vec{d}] = \nabla f(\vec{x}) \cdot \vec{d}$ for all $d \in \mathbb{R}^n$, allowing us to separate the direction $\vec{d}$ and the actual multidimensional derivative.

Gradient

It turns out, assuming that the function f is differentiable, that there exists a vector ∇f such that $\nabla f(\vec{x})[\vec{d}] = \nabla f(\vec{x}) \cdot \vec{d}$ for all $d \in \mathbb{R}^n$, allowing us to separate the direction $\vec{d}$ and the actual multidimensional derivative.

In particular, the expression for this $\nabla f(\vec{x})$ that satisfies the above property is

$$\nabla f(\vec{x}) = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix}.$$

This gets the name "gradient" as it represents the set of slopes around a point as one moves one unit in each dimension parallel to the n axes.

Gradient

Let's formally define the gradient.

Definition

The **gradient** vector represents the derivative for a function $f : \mathbb{R}^n \to \mathbb{R}$. If the function is differentiable, the gradient is equal to the $1 \times n$ vector where the ith entry is $\left[\frac{\partial f}{\partial \vec{x}}\right]_{\text{i}} = \frac{\partial f}{\partial x_i}$.

Gradient

Let's formally define the gradient.

Definition

The **gradient** vector represents the derivative for a function $f : \mathbb{R}^n \to \mathbb{R}$. If the function is differentiable, the gradient is equal to the $1 \times n$ vector where the ith entry is $\left[\frac{\partial f}{\partial \vec{x}}\right]_{\mathrm{i}} = \frac{\partial f}{\partial x_i}$.

Example

Consider the function $f(\vec{x}) = (x_1^2 + x_2^2)$. Then $\frac{\partial f}{\partial x_1} = 2x_1, \frac{\partial f}{\partial x_2} = 2x_2$, and so $\frac{df}{d\vec{x}} = \nabla f(\vec{x}) = (2x_1, 2x_2)$.

Jacobian

Next, we turn to functions $\vec{f} : \mathbb{R}^n \to \mathbb{R}^m$ where both the input and output are vectors. We treat the gradient vectors for each entry separately.

As we have defined the gradient for a single-variable function as a row vector, for a function with vector output we could stack these m row vectors on top of one another to get an $m \times n$ matrix.

This matrix is called the Jacobian.

$$\begin{bmatrix} \frac{df_1}{d\vec{x}} \\ \frac{df_2}{d\vec{x}} \\ \vdots \\ \frac{df_m}{d\vec{x}} \end{bmatrix} = \begin{bmatrix} \nabla f_1(\vec{x}) \\ \nabla f_2(\vec{x}) \\ \vdots \\ \nabla f_m(\vec{x}) \end{bmatrix} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix}.$$

Jacobian

This definition allows us to extend the limit definition of a multivariable derivative to the Jacobian, as it only involves stacking gradients:

$$\lim_{t \to 0} \frac{\vec{f}(\vec{x} + t\vec{d}) - \vec{f}(\vec{x})}{t} = \nabla \vec{f}(\vec{x})[\vec{d}\,] = \begin{bmatrix} \nabla f_1(\vec{x}) \\ \nabla f_2(\vec{x}) \\ \vdots \\ \nabla f_n(\vec{x}) \end{bmatrix} \cdot \vec{d} = J_f \cdot \vec{d}.$$

Jacobian

Definition

The **Jacobian** matrix represents the derivative for a function $f : \mathbb{R}^n \to \mathbb{R}^m$. It is defined as the $m \times n$ matrix where the term at the ith row and jth column is $\left[\frac{\partial \vec{f}}{\partial \vec{x}}\right]_{\mathrm{ij}} = \frac{\partial f_i}{\partial x_j}$.

Example

Consider the function $\vec{f}(\vec{x}) = (x_1^2 + x_2^2, x_2^3)$. Then,

$$\frac{d\vec{f}}{d\vec{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 & 2x_2 \\ 0 & 3x_2^2 \end{bmatrix}$$

.

Hessian

In the one-dimensional case, $f : \mathbb{R} \to \mathbb{R}$, we can take the derivative $f'(x)$, and then take the derivative $f''(x)$ of $f'(x)$.

Now, if we have $f : \mathbb{R}^n \to \mathbb{R}$, the cross-partial derivatives (i.e. $\frac{\partial^2 f}{\partial x_i \partial x_j}$ may be of interest in certain applications.

We can arrange this into a matrix, called the Hessian matrix $\mathbf{H}_f$ of function f, such that the entry at the ith row and jth column of the Hessian is the partial derivative with respect to x_i of the partial derivative of the function with respect to x_j.

Hessian

This definition of $\mathbf{H}_f$ gives an analogue of the second derivative for vectors, where the second derivative is taken with respect to the transpose of the vector.

Indeed, basing off the above definition, we notice that the ith column of the Hessian $\mathbf{H}_f$ is the the gradient vector of the ith entry of the gradient $\nabla f(\vec{x})$, with respect to $\vec{x}$.

Based on this interpretation, if we consider the gradient to be a function itself, as in $\nabla f : \mathbb{R}^n \to \mathbb{R}^n$, transpose into a column vector, and then taking the Jacobian of the transpose.

Transposing again gives the Hessian H_f:

$$\mathbf{H}_f(\vec{x}) = \mathbf{J}_f((\nabla f(\vec{x}))^T)^T$$

.

Hessian

Based on the above definition, we thus can write:

$$\mathbf{H}_f(\vec{x}) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix} = \begin{bmatrix} \frac{\partial}{\partial \vec{x}}[\nabla f(\vec{x})]_1 \\ \frac{\partial}{\partial \vec{x}}[\nabla f(\vec{x})]_2 \\ \vdots \\ \frac{\partial}{\partial \vec{x}}[\nabla f(\vec{x})]_n \end{bmatrix}^T = J_f((\nabla f(\vec{x})^{\mathbf{T}}))^T$$

Definition

The *Hessian* matrix, denoted $\mathbf{H}_f$ represents the second derivative for a function $f : \mathbb{R}^n \to \mathbb{R}$. It is defined as the $n \times n$ matrix where the term at the ith row and jth column is $\left[\frac{\partial^2 f}{\partial \vec{x} \partial \vec{x}^{\mathsf{T}}}\right]_{\mathrm{ij}} = \frac{\partial^2 f}{\partial x_i \partial x_j}$.

Hessian

Based on the above definition, we thus can write:

$$\mathbf{H}_f(\vec{x}) = \begin{bmatrix} \frac{\partial^2 f(x)}{\partial x_1^2} & \frac{\partial^2 f(x)}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_1 \partial x_n} \\ \frac{\partial^2 f(x)}{\partial x_2 \partial x_1} & \frac{\partial^2 f(x)}{\partial x_2^2} & \cdots & \frac{\partial^2 f(x)}{\partial x_2 \partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f(x)}{\partial x_n \partial x_1} & \frac{\partial^2 f(x)}{\partial x_n \partial x_2} & \cdots & \frac{\partial^2 f(x)}{\partial x_n^2} \end{bmatrix} = \begin{bmatrix} \dfrac{\partial}{\partial \vec{x}}[\nabla f(\vec{x})]_1 \\ \dfrac{\partial}{\partial \vec{x}}[\nabla f(\vec{x})]_2 \\ \vdots \\ \dfrac{\partial}{\partial \vec{x}}[\nabla f(\vec{x})]_n \end{bmatrix}^T = J_f((\nabla f(\vec{x}))^T$$

The Hessian matrix is symmetric because $\dfrac{\partial^2 f}{\partial x_i \partial x_j} = \dfrac{\partial^2 f}{\partial x_j \partial x_i}$ subject to certain analytic conditions that are satisfied by most continuous functions used in statistics.

Hessian

Example

Let $f(\vec{x}) = x_2^2 + x_1 x_2 + x_3^2$. Then

$$\frac{\partial}{\partial x_1}\left[\frac{\partial}{\partial x_2} f(\vec{x})\right] = \frac{\partial}{\partial x_1}(2x_2 + x_1) = 1.$$

Another example is that

$$\frac{\partial}{\partial x_3}\left[\frac{\partial}{\partial x_3}\right] = \frac{\partial}{\partial x_3}(2x_3) = 2.$$

Computing all the other entries, we can get that the Hessian is

$$\begin{bmatrix} 0 & 1 & 0 \\ 1 & 2 & 0 \\ 0 & 0 & 2 \end{bmatrix}.$$

Motivation for Layouts

At this point, although the definitions for the gradient, Jacobian, and Hessian may make intuitive sense, some aspects of our definition may seem to be a matter of preference.

For example, in the definition of the gradient, it is not clear why it is defined as a row rather than a column of partial derivatives.

Motivation for Layouts

At this point, although the definitions for the gradient, Jacobian, and Hessian may make intuitive sense, some aspects of our definition may seem to be a matter of preference.

For example, in the definition of the gradient, it is not clear why it is defined as a row rather than a column of partial derivatives.

Consider $f : \mathbb{R}^n \to \mathbb{R}^m$. Generalizing from the limit definition of a derivative, we could write the linear approximation form

$$\Delta\vec{f} = \left(\frac{d\vec{f}}{d\vec{x}}\right) \cdot \Delta\vec{x}.$$

As $\Delta\vec{f}$ is an $m \times 1$ vector and $\Delta\vec{x}$ is an $n \times 1$ vector, a single expression for $\frac{d\vec{f}}{d\vec{x}}$ will be an $m \times n$ matrix.

Motivation for Layouts

In the scalar case, this definition is the only sensible one. In the case of a scalar-valued function f, however, we might want to make Δf be the dot product of $\frac{df}{d\vec{x}}$ and $\Delta\vec{x}$, as this is typically how vectors are multiplied.

This gives the equation

$$\Delta f = \left(\frac{df}{d\vec{x}}\right)^T \cdot \Delta\vec{x},$$

and extending to a vector-valued function f gives

$$\Delta\vec{f} = \left(\frac{d\vec{f}}{d\vec{x}}\right)^T \cdot \Delta\vec{x}.$$

As $\left(\frac{d\vec{f}}{d\vec{x}}\right)^T$ has dimension $m \times n$, $\left(\frac{d\vec{f}}{d\vec{x}}\right)$ has dimension $n \times m$.

Motivation for Layouts

The two possible equations representing the same concept is the basis for the two multivariable differentiation layouts.

- $\Delta \vec{f} = \left(\frac{d\vec{f}}{d\vec{x}} \right) \cdot \Delta \vec{x}$ leads to the numerator layout
- $\Delta f = \left(\frac{df}{d\vec{x}} \right)^T \cdot \Delta \vec{x}$ leads to the denominator layout

Motivation for Layouts

The two possible equations representing the same concept is the basis for the two multivariable differentiation layouts.

- $\Delta\vec{f} = \left(\frac{d\vec{f}}{d\vec{x}}\right) \cdot \Delta\vec{x}$ leads to the numerator layout
- $\Delta f = \left(\frac{df}{d\vec{x}}\right)^T \cdot \Delta\vec{x}$ leads to the denominator layout

There is no one correct layout as it is possible to derive internally consistent composition rules for both layouts.

Our composition rules below do *not* work even when you use denominator layout consistently!

Numerator Layout

If we start with the equation $\Delta\vec{f} = (\frac{d\vec{f}}{d\vec{x}}) \cdot \Delta\vec{x}$, then column i of the matrix $\frac{d\vec{f}}{d\vec{x}}$ must correspond to ith entry of the vector $\Delta\vec{x}$ for the dot product to hold, i.e.

$$\Delta\vec{f} = \sum_{i=1}^{n} \frac{\partial\vec{f}}{\partial x_i}\Delta x_i.$$

Hence each $\frac{\partial\vec{f}}{\partial x_i}$ would be the ith column of the matrix $\frac{d\vec{f}}{d\vec{x}}$, yielding

$$\frac{d\vec{f}}{d\vec{x}} = \begin{bmatrix} \frac{\partial\vec{f}}{\partial x_1} & \frac{\partial\vec{f}}{\partial x_2} & \dots & \frac{\partial\vec{f}}{\partial x_n} \end{bmatrix}.$$

Numerator Layout

Expanding each of the columns along the components of $\vec{f}$ gives

$$\frac{d\vec{f}}{d\vec{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_2}{\partial x_n} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_m}{\partial x_1} & \frac{\partial f_m}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} = \left[\frac{\partial f_i}{\partial x_j}\right]_{1\le i\le m, 1\le j\le n}.$$

Numerator Layout

Another view involves compressing each row of the expanded form as a single expression, as in

$$\frac{d\vec{f}}{d\vec{x}} = \begin{bmatrix} \frac{df_1}{d\vec{x}} \\ \frac{df_2}{d\vec{x}} \\ \cdots \\ \frac{df_m}{d\vec{x}} \end{bmatrix}.$$

This is called the **numerator layout** because the expression in the numerator, which is a column vector, maintains its orientation in the matrix $\frac{d\vec{f}}{d\vec{x}}$, if the expression is specialized to a single x_i. Conversely, the expression in the denominator, which is also a column vector, has its orientation transposed in $\frac{d\vec{f}}{d\vec{x}}$, if the expression is specialized to a single f_j.

Denominator Layout

On the other hand, if we start with the equation $\Delta\vec{f} = (\frac{d\vec{f}}{d\vec{x}})^T \cdot \Delta\vec{x}$, then row i of the matrix $\frac{d\vec{f}}{d\vec{x}}$ must correspond to ith entry of the vector $\Delta\vec{x}$ for the dot product to hold, i.e.

$$\Delta\vec{f} = \sum_{i=1}^{n}(\frac{\partial\vec{f}}{\partial x_i})^T \Delta x_i.$$

Denominator Layout

Hence each $\frac{\partial \vec{f}}{\partial x_i}$ would be the ith row of the matrix $\frac{d\vec{f}}{d\vec{x}}$, yielding

$$\frac{d\vec{f}}{d\vec{x}} = \begin{bmatrix} \frac{\partial \vec{f}}{\partial x_1} \\ \frac{\partial \vec{f}}{\partial x_2} \\ \cdots \\ \frac{\partial \vec{f}}{\partial x_n} \end{bmatrix}.$$

Denominator Layout

Expanding each of the columns along the components of $\vec{f}$ gives

$$\frac{d\vec{f}}{d\vec{x}} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_2}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_1} \\ \frac{\partial f_1}{\partial x_2} & \frac{\partial f_2}{\partial x_2} & \cdots & \frac{\partial f_m}{\partial x_2} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f_1}{\partial x_n} & \frac{\partial f_2}{\partial x_n} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} = \left[\frac{\partial f_j}{\partial x_i}\right]_{1 \le i \le n, 1 \le j \le m}.$$

Denominator Layout

Another view involves compressing each column of the expanded form as a single expression, as in

$$\frac{d\vec{f}}{d\vec{x}} = \begin{bmatrix} \frac{df_1}{d\vec{x}} & \frac{df_2}{d\vec{x}} & \cdots & \frac{df_m}{d\vec{x}} \end{bmatrix}.$$

This is called the **denominator layout** because the expression in the denominator, which is a column vector, maintains its orientation in the matrix $\frac{d\vec{f}}{d\vec{x}}$, if the expression is specialized to a single f_i.

Conversely, the expression in the numerator, which is also a column vector, has its orientation transposed in $\frac{d\vec{f}}{d\vec{x}}$, if the expression is specialized to a single x_j. *Note that this class does not use this layout, we simply present this for informational reasons.*

Introduction

The previous sections have discussed how to differentiate functions of the general form $f : \mathbb{R}^n \to \mathbb{R}^m$ and take the second derivatives for functions of the form $f : \mathbb{R}^n \to \mathbb{R}$.

If we allow either the numerator or denominator expression in our derivative to go beyond vectors, however, we run into a serious dimension constraint, as a matrix has only two dimensions.

Therefore, we will only consider two forms of matrix derivatives:

- $\dfrac{df}{d\mathbf{X}}$ $(f : \mathbb{R}^{m \times n} \to \mathbb{R})$, and
- $\dfrac{d\mathbf{F}}{dx}$ $(\mathbf{F} : \mathbb{R} \to \mathbb{R}^{m \times n})$.

$\frac{df}{d\mathbf{X}}$

This matrix derivative represents the matrix analogue for a gradient.

In the numerator layout, treating $\mathbf{X}$ as a row vector of column vectors, $\frac{df}{d\mathbf{X}}$ has dimension $n \times m$ and could be seen as column vector of the gradients of f with respect to each of the column vectors in $\mathbf{X}$.

$$\frac{df}{d\mathbf{X}} = \begin{bmatrix} \frac{\partial f}{\partial \vec{\mathbf{X}}_{\cdot 1}} \\ \frac{\partial f}{\partial \vec{\mathbf{X}}_{\cdot 2}} \\ \vdots \\ \frac{\partial f}{\partial \vec{\mathbf{X}}_{\cdot n}} \end{bmatrix} = \begin{bmatrix} \frac{\partial f}{\partial \mathbf{X}_{11}} & \frac{\partial f}{\partial \mathbf{X}_{21}} & \cdots & \frac{\partial f}{\partial \mathbf{X}_{m1}} \\ \frac{\partial f}{\partial \mathbf{X}_{12}} & \frac{\partial f}{\partial \mathbf{X}_{22}} & \cdots & \frac{\partial f}{\partial \mathbf{X}_{m2}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial \mathbf{X}_{1n}} & \frac{\partial f}{\partial \mathbf{X}_{2n}} & \cdots & \frac{\partial f}{\partial \mathbf{X}_{mn}} \end{bmatrix} = \left[\frac{\partial f}{\partial \mathbf{X}_{ji}}\right]_{1 \le i \le m, 1 \le j \le n}.$$

$\frac{d\mathbf{F}}{dx}$

This matrix derivative represents the matrix analogue for a parametrized vector derivative.

In the numerator layout, $\frac{d\mathbf{F}}{dx}$ is an $m \times n$ matrix where each entry is the derivative of the corresponding entry in $\mathbf{F}$ with respect to x.

$$\frac{d\mathbf{F}}{dx} = \begin{bmatrix} \frac{d\mathbf{F}_{11}}{dx} & \frac{d\mathbf{F}_{12}}{dx} & \cdots & \frac{d\mathbf{F}_{1n}}{dx} \\ \frac{d\mathbf{F}_{21}}{dx} & \frac{d\mathbf{F}_{22}}{dx} & \cdots & \frac{d\mathbf{F}_{2n}}{dx} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{d\mathbf{F}_{m1}}{dx} & \frac{d\mathbf{F}_{m2}}{dx} & \cdots & \frac{d\mathbf{F}_{mn}}{dx} \end{bmatrix} = \left[\frac{d\mathbf{F}_{ij}}{dx}\right]_{1\leq i\leq m, 1\leq j\leq n}.$$

Basic Example

Example

For any n-dimensional vector $\vec{x}$,

$$\frac{d\vec{x}}{d\vec{x}} = \mathbf{I}_n,$$

where $\mathbf{I}_n$ is the $n \times n$ identity matrix.

Proof: By definition, for any n-dimensional vector $\vec{d}$,

$$\begin{aligned}\frac{d\vec{x}}{d\vec{x}} \cdot \vec{d} &= \lim_{t \to 0} \frac{(\vec{x} + t\vec{d}) - (\vec{x})}{t} \\ &= \vec{d}.\end{aligned}$$

We note that $\frac{d\vec{x}}{d\vec{x}} = \mathbf{I}_n$ satisfies the limit definition.

More Involved Example

Example

For any n-dimensional vector $\vec{x}$ and $n \times n$ constant matrix A,

$$\frac{d\vec{x}^T \mathbf{A}\vec{x}}{d\vec{x}} = \vec{x}^T(\mathbf{A} + \mathbf{A}^T).$$

Proof: By definition, for any n-dimensional vector $\vec{d}$,

$$\begin{aligned}\frac{d\vec{x}^T \mathbf{A}\vec{x}}{d\vec{x}} \cdot \vec{d} &= \lim_{t\to 0} \frac{(\vec{x}+t\vec{d})^T \mathbf{A}(\vec{x}+t\vec{d}) - \vec{x}^T\mathbf{A}\vec{x}}{t} \\ &= \lim_{t\to 0} \left(\vec{d}^T\mathbf{A}\vec{x} + \vec{x}^T\mathbf{A}\vec{d} + t\vec{d}^T\mathbf{A}\vec{d}\right) \\ &= \vec{d}^T\mathbf{A}\vec{x} + \vec{x}^T\mathbf{A}\vec{d} \\ &= \vec{x}^T\mathbf{A}^T\vec{d} + \vec{x}^T\mathbf{A}\vec{d} \\ &= \vec{x}^T(\mathbf{A}+\mathbf{A}^T)\vec{d}.\end{aligned}$$

Constant Product

Since left- or right-multiplying by a matrix (or vector or scalar) are linear transformations, the linearity of differentiation allows any constant factors to be factored out of the differential term:

1. $\dfrac{d(au)}{d\mathbf{X}} = a\dfrac{du}{d\mathbf{X}}$ if a is constant with respect to $\mathbf{X}$.
2. $\dfrac{d\mathbf{A}\vec{u}}{d\vec{x}} = \mathbf{A}\dfrac{d\vec{u}}{d\vec{x}}$ if $\mathbf{A}$ is a constant w.r.t $\vec{x}$.
3. $\dfrac{d\mathbf{AUB}}{dx} = \mathbf{A}\dfrac{d\mathbf{U}}{dx}\mathbf{B}$ if $\mathbf{A}, \mathbf{B}$ are constants with respect to x.

Entry-based proofs of these formulas are left as an exercise.

Addition Rule

The linearity of differentiation shows the derivative of a sum is the sum of the derivatives of the summands:

1. $$\frac{d(u+v)}{d\mathbf{X}} = \frac{du}{d\mathbf{X}} + \frac{dv}{d\mathbf{X}}$$
2. $$\frac{d(\vec{u}+\vec{v})}{d\vec{x}} = \frac{d\vec{u}}{d\vec{x}} + \frac{d\vec{v}}{d\vec{x}}$$
3. $$\frac{d(\mathbf{U}+\mathbf{V})}{dx} = \frac{d\mathbf{U}}{dx} + \frac{d\mathbf{V}}{dx}$$

Entry-based proofs of these formulas are left as an exercise.

Product Rule

We know that a product rule for vector and matrix differentiation necessarily holds for one-entry vectors and matrices, so it necessarily has a similar form as the product rule for scalars. However, bear in mind that transposes and the order of multiplication both matter when the terms involved are multidimensional.

Product Rule

We know that a product rule for vector and matrix differentiation necessarily holds for one-entry vectors and matrices, so it necessarily has a similar form as the product rule for scalars. However, bear in mind that transposes and the order of multiplication both matter when the terms involved are multidimensional.

1. $$\frac{duv}{d\mathbf{X}} = u\frac{dv}{d\mathbf{X}} + v\frac{du}{d\mathbf{X}}$$
2. $$\frac{d\vec{u}^T\vec{v}}{d\vec{x}} = \vec{u}^T\frac{d\vec{v}}{d\vec{x}} + \vec{v}^T\frac{d\vec{u}}{d\vec{x}}$$
3. $$\frac{d\mathbf{UV}}{dx} = \mathbf{U}\frac{d\mathbf{V}}{dx} + \frac{d\mathbf{U}}{dx}\mathbf{V}$$

We will use an entry-wise proof for Rule 3; similar proofs for the other formulae are left as an exercise.

Proof of $\frac{d\mathbf{UV}}{dx} = \mathbf{U}\frac{d\mathbf{V}}{dx} + \frac{d\mathbf{U}}{dx}\mathbf{V}$

Proof.

For any indices i, j,

$$\begin{aligned}
\left[\frac{d\mathbf{UV}}{dx}\right]_{ij} = \frac{d[\mathbf{UV}]_{ij}}{dx} &= \sum_k \frac{d\mathbf{U}_{ik}\mathbf{V}_{kj}}{dx} \\
&= \sum_k \left(\mathbf{U}_{ik}\frac{d\mathbf{V}_{kj}}{dx} + \mathbf{V}_{kj}\frac{d\mathbf{U}_{ik}}{dx}\right) \\
&= \sum_k \mathbf{U}_{ik}\left[\frac{d\mathbf{V}}{dx}\right]_{kj} + \sum_k \mathbf{V}_{kj}\left[\frac{d\mathbf{U}}{dx}\right]_{ik} \\
&= \left[\mathbf{U}\frac{d\mathbf{V}}{dx}\right]_{ij} + \left[\frac{d\mathbf{U}}{dx}\mathbf{V}\right]_{ij}.
\end{aligned}$$

□

Chain Rule

Similar to the one-dimensional chain rule, we can write the derivative of y with respect to x when $y = y(u)$, $u = u(x)$ in terms of $\frac{dy}{du}$ and $\frac{du}{dx}$.

However, in the multidimensional case, we instead seek is the *sum* of the partial derivatives; in particular, if u_i is a (scalar) entry of u, then the total derivative is

$$\frac{dy}{dx} = \sum_i \frac{\partial y}{\partial u_i} \frac{du_i}{dx}$$

1. $\frac{dy}{d\mathbf{X}} = \frac{dy}{du} \cdot \frac{du}{d\mathbf{X}}$
2. $\frac{d\vec{y}}{d\vec{x}} = \frac{d\vec{y}}{d\vec{u}} \cdot \frac{d\vec{u}}{d\vec{x}}$
3. $\frac{d\mathbf{Y}}{dx} = \frac{d\mathbf{Y}}{du} \cdot \frac{du}{dx}$

Numerator layout only

Format reversed in denominator layout

Proof of $\frac{d\vec{y}}{d\vec{x}} = \frac{d\vec{y}}{d\vec{u}} \cdot \frac{d\vec{u}}{d\vec{x}}$

Proof.

For any indices i, j,

$$\begin{aligned}\left[\frac{d\vec{y}}{d\vec{x}}\right]_{ij} &= \frac{\partial y_i}{\partial x_j} \\ &= \sum_k \frac{\partial y_i}{\partial u_k}\frac{\partial u_k}{\partial x_j} \\ &= \sum_k \left[\frac{d\vec{y}}{d\vec{u}}\right]_{ik}\left[\frac{d\vec{u}}{d\vec{x}}\right]_{kj} \\ &= \left[\frac{d\vec{y}}{d\vec{u}} \cdot \frac{d\vec{u}}{d\vec{x}}\right]_{ij}.\end{aligned}$$

Hence $\frac{d\vec{y}}{d\vec{x}} = \frac{d\vec{y}}{d\vec{u}} \cdot \frac{d\vec{u}}{d\vec{x}}$ as desired. □

Chain Rule

We note that we cannot use a matrix as $\mathbf{U}$ in a chain rule with our definitions so far.

Chain Rule

We note that we cannot use a matrix as $\mathbf{U}$ in a chain rule with our definitions so far.

The reason that we can use a vector as $\vec{u}$ is because matrix multiplication sums up the partial derivatives correctly; each row in $\frac{d\vec{y}}{d\vec{u}}$ is the row vector $\frac{d\vec{y}_i}{d\vec{u}}$, and each column in $\frac{d\vec{u}}{d\vec{x}}$ is the column vector $\frac{d\vec{u}}{d\vec{x}_j}$. Matrix multiplication takes the dot product of the two vectors, thus summing the product of the corresponding partial derivatives together.

We can easily observe that if $\mathbf{U}$ were a matrix and x, y scalars, the matrix product would not sum up $\frac{\partial y}{\partial \mathbf{U}_{ij}} \cdot \frac{\partial \mathbf{U}_{ij}}{\partial x}$ correctly. The correct approach is to take the sum of the element-wise product of the two matrices.