Matrix Calculus for 10-301/601

Hoeseong (Hayden) Kim
Abhishek Vijayakumar

Carnegie Mellon University

February 24, 2022

How to Read

Please read this first!

What is this write-up?

This write-up covers everything you need to know (and a little more) about matrix calculus to pass 10-301/601. You must be fairly comfortable with single-variable calculus and basic vector algebra before reading this (and for 10-301/601). This does not constitute as a formal introduction to matrix calculus, but anything necessary for the course is covered.

What topics are covered in this write-up, and when should I read this?

The first section glosses over basic multivariable calculus you need for the class, such as gradients and partial derivatives. You may skip this section if you are already familiar with this topic, but please do not skip the first exercise question. Topics in this section will be covered in the first exam, so it is highly recommended that you read this as early as possible.

The second section introduces basic definitions of matrix derivatives and how the chain rule is extended to matrix calculus. You do not need any prior knowledge on deep learning. Aim to fully understand this section before the release of homework 5. This will help you greatly with the chain rule and back propagation part of the course.

The last section focuses more on how to actually compute the derivatives (who uses the definition of the derivative to find the derivative of $y = 3x^2 + 5$?). You will learn to use how to derive different versions of chain rules, and how to compute any derivatives you will encounter in 10-301/601 starting from considering one element of the result. This section will be the most helpful section for the homework and exams.

How should I solve the exercises?

Each section includes exercises that help you understand or apply the material. ***Do NOT skip the exercises***, as they also introduce some new theorems and facts that are greatly useful for the course. Practice makes perfect, especially for math! The exercises are designed to be solved (mostly) in order. Some of them may depend on the results derived in previous exercises.

When/How should I read the solutions?

All exercises are accompanied with fairly detailed solutions, especially for Sections 2 and 3. Avoid reading the solutions before properly attempting to solve the problems. When you are stuck, read the section again, digest the content, and come back to it later; maybe collaborate with others if necessary. Please do not resort to the solutions before giving yourself enough time to think about the question.

Make sure to compare your solutions with the reference solutions. Some questions have multiple solutions with different approaches, from which you may be able to develop more intuition. If you find any errors or have a better/more efficient solution or any feedback, please send me an email!

Contents

1 Multivariable Scalar Functions

This section briefly summarizes some important concepts of multivariable calculus. We will skip any mathematical details or proofs not necessary for the course. Some important concepts such as the definition of limit, continuity, differentiability are omitted since they are not the focus of 10-301/601, but they are not to be made light of.

1.1 $\mathbb{R}^n \to \mathbb{R}$ Functions

In this section, we deal with functions that map a vector $\mathbb{R}^n$ to a scalar $\mathbb{R}$. We use *column vectors* by default throughout the entire write-up.* Such $\mathbb{R}^n \to \mathbb{R}$ functions can also be considered to take multiple scalar inputs and yield one scalar output. Some examples include:

1. The volume of a cone whose radius of the base is r and the height is h is given as:

$$V(r, h) = \frac{1}{3}\pi r^2 h.$$

   The function V maps a vector $[r, h]^T \in \mathbb{R}^2$ to a scalar $\frac{1}{3}\pi r^2 h \in \mathbb{R}$.

2. The distance between two points a and b on the x-axis is given as:

$$d(a, b) = |a - b|.$$

   The function d maps a vector $[a, b]^T \in \mathbb{R}^2$ to a scalar $|a - b| \in \mathbb{R}$.

3. *(Important)* The L_2 norm of a vector $\mathbf{x} = [x_1, x_2, \cdots, x_n]^T \in \mathbb{R}^n$ is given as:

$$f(\mathbf{x}) = \|\mathbf{x}\|_2 = \|\mathbf{x}\| = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}.^\dagger$$

   The function f maps a vector $\mathbf{x} \in \mathbb{R}^n$ to a scalar $\sqrt{x_1^2 + \cdots + x_n^2} \in \mathbb{R}$. This example is marked as important because you will use L_2 norm a lot, and because you will often see a vector itself being passed to a function. This can be thought of as the following:

$$f(x_1, x_2, \cdots, x_n) = \sqrt{x_1^2 + x_2^2 + \cdots + x_n^2}.$$

1.2 Partial Derivatives

Recall how we took the derivative of a $\mathbb{R} \to \mathbb{R}$ function. A simple function, say $f(x) = x^2$, has only one independent variable x, and naturally we take the derivative of x^2 with respect to that independent variable, x. The key point here is that there is only *one* input, so we have no other choice but to differentiate with respect to that one variable. Now for $\mathbb{R}^n \to \mathbb{R}$ functions, we have n inputs, so we end up with more possible choices—with respect to which variable do we differentiate f?

*The write-up follows the convention used in class. More about the notation can be found here.

†Note that the subscript 2 can be omitted for L_2 norm.

The derivative with respect to a single independent variable is obtained by simply pretending as if all the other variables are constants. For example, consider

$$f(x, y, z) = xy + y^x + 2z$$

and say we are taking the derivative with respect to one of the variables, y. Then we treat x and z as constants, and the result will be:

$$x + xy^{x-1}.$$

We walk through this result term by term. For xy, only y is regarded as a variable and x is considered as a constant, so the derivative is x. This is analogous to the derivative of $3x$ being 3; x is a variable and 3 is a constant. For y^x, again, x is treated as a constant so we have xy^{x-1} (just like how $(x^3)' = 3x^2$). For $2z$, the entire term is a constant and the derivative is zero. We call what we just evaluated a **partial derivative** of f with respect to y, and mathematically we write:

$$\frac{\partial f}{\partial y} = x + xy^{x-1}, \quad \text{or}$$
$$\nabla_y f(x, y) = x + xy^{x-1}.$$

The symbol ∂ is read "partial," and ∇ is read "nabla," "del," or "gradient."

Just as we can differentiate a single-variable function multiple times, we may be interested in evaluating higher order partial derivatives. Recall that higher order derivatives are written as:

$$\frac{d^2 f}{dx^2}, \frac{d^3 f}{dx^3}, \cdots, \frac{d^n f}{dx^n}.$$

Similarly, when we take the partial derivative multiple times with respect to the same variable, we write:

$$\frac{\partial^2 f}{\partial x^2}, \frac{\partial^3 f}{\partial x^3}, \cdots, \frac{\partial^n f}{\partial x^n}.$$

However, because now we have multiple input variables, we do not necessarily have to take the partial derivative with respect to the same variable every time. For $f(x, y, z) = xy + y^x + 2z$, we can take the partial derivative with respect to y and then z. This is written as

$$\frac{\partial^2 f}{\partial z \partial y} = \frac{\partial}{\partial z}\left[x + xy^{x-1}\right] = 0.$$

The power of the "numerator" means how many times we differentiate, and the "denominator" determines which variables we take the partial derivatives with respect to and in what order. Remember that you have to read it *right-to-left*; $\partial z \partial y$ means with respect to y first, not z! It is worth mentioning that you can change the order in which partial derivatives are taken under certain conditions, i.e.,

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}.$$

A lot of the functions we will encounter have this property. This, however, is not true in general.*

*This holds when the partial derivatives exist and are continuous in an open region containing the point at which the partial derivative is evaluated. In 10-301/601, this is almost always the case.

1.3 Gradients

Instead of having to inspect the partial derivatives one by one, what if we want a single entity that represents the degree of change with respect to all variables altogether? This motivates the use of **gradient**, which is simply a vector of all partial derivatives. For example, for $f(x, y, z) = xy + y^x + 2z$, the gradient is:

$$\begin{bmatrix} \partial f/\partial x \\ \partial f/\partial y \\ \partial f/\partial z \end{bmatrix} = \begin{bmatrix} y + y^x \log y \\ x + xy^{x-1} \\ 2 \end{bmatrix}.$$

Mathematically, we write:

$$\nabla f(x) = \begin{bmatrix} y + y^x \log y \\ x + xy^{x-1} \\ 2 \end{bmatrix}.$$

You may see ∇ in boldface or with an arrow on top to emphasize that it is a vector.

Gradient is extremely important and utilized a lot in machine learning. One of the most important properties of gradient is that the gradient of a function evaluated at one point is the direction to take in order to climb up the function the fastest. In other words, the exact opposite direction of the gradient vector is the direction to take to climb down the function the fastest (Figure 1).

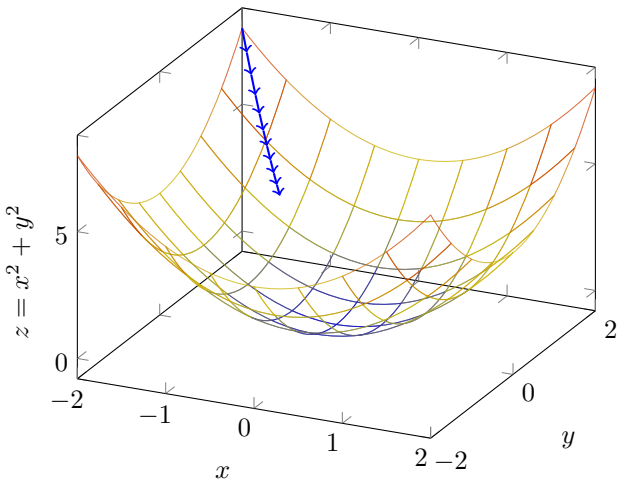


Figure 1: Climbing down $z = x^2 + y^2$ from point $(-2, 2, 8)$ following the *opposite* direction of the gradient vector.

1.4 Exercises

1. In this problem, we will briefly review single-variable calculus with some extremely useful functions for deep learning.

   (a) Evaluate $\frac{d}{dx}\sigma(x)$ where $\sigma(x) = 1/(1+e^{-x})$. This is called the sigmoid function.

   (b) Express your answer in (a) using only $\sigma(x)$ and constants.

   (c) Evaluate $\frac{d}{dx}\tanh(x)$ where $\tanh(x) = (e^x - e^{-x})/(e^x + e^{-x})$. This is called the hyperbolic tangent function.

   (d) Express your answer in (c) using only $\tanh(x)$ and constants.

2. Evaluate the following:

   (a) $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ where $f(x,y) = x^y + y^x$

   (b) $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ where $f(x,y) = \sin(y + \cos x)$

   (c) $\frac{\partial f}{\partial x}$ and $\frac{\partial f}{\partial y}$ where $f(x,y) = e^{xy} + y\log 3x$

   (d) $\frac{\partial^2 f}{\partial x^2}$, $\frac{\partial^2 f}{\partial x \partial y}$, $\frac{\partial^2 f}{\partial y \partial x}$, and $\frac{\partial^2 f}{\partial y^2}$ where $f(x,y) = \sin(xy) + \cos(xy)$

   (e) $\nabla_x f(x,y)$ and $\nabla_y f(x,y)$ where $f(x,y) = x^{\log y} + x^2 + 2y$

   (f) $\nabla_x f(x,y)$ and $\nabla_y f(x,y)$ where $f(x,y) = (x+y)^2$

   (g) $\frac{\partial f}{\partial x_i}$ where $f(\mathbf{x}) = \|\mathbf{x}\|_2^2$ $(1 \le i \le n)$ *Hint:* Recall that $\|\mathbf{x}\|_2 = \sqrt{x_1^2 + \cdots + x_n^2}$

   (h) $\frac{\partial f}{\partial x_i}$ where $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ and $\mathbf{w}$ is a constant vector $(1 \le i \le n)$

3. Evaluate the following:

   (a) $\nabla f(x,y)$ where $f(x,y) = xy^2 + x^2y$

   (b) $\nabla f(x,y)$ where $f(x,y) = (x+y)^2$

   (c) $\nabla^2 f(x,y)$ where $f(x,y) = \sin(e^{xy})$

   (d) $\nabla f(\mathbf{x})$ where $f(\mathbf{x}) = \|\mathbf{x}\|_2^2$

   (e) Express your answer in (d) using only one variable (no limit on constants).

   (f) $\nabla f(\mathbf{x})$ where $f(\mathbf{x}) = \mathbf{w}^T\mathbf{x}$ and $\mathbf{w}$ is a constant vector

   (g) Express your answer in (f) using only one variable (no limit on constants).

4. Hayden was taking a nap on a hill at Schenley park, only to realize that he has to run back to the campus for his next class in two minutes. He approximates the height h of the hill at position (x,y) as $h = x^2 - 3y^2$, and guesses that his current position is $(x,y,h) = (-1,0,1)$. Which direction should he take to go down the hill as fast as possible?

2 Basics of Matrix Calculus

In this section, we will cover the basic definitions of matrix calculus and how the chain rule works in matrix calculus.

2.1 Definitions

In the world of single-variable functions, the options are limited for taking the derivative; for $f : \mathbb{R} \to \mathbb{R}, x \mapsto f(x)$, the only derivative of our interest is $\frac{df}{dx}$. But with functions such as $g(\mathbf{x}) = \mathbf{A}\mathbf{x}$ and $h(\mathbf{x}, \mathbf{A}) = \mathbf{x}^T\mathbf{A}\mathbf{x}$, we can also consider derivatives such as $\frac{dg}{d\mathbf{x}}$, $\frac{dg}{dx_i}$, $\frac{dh}{d\mathbf{A}}$, $\frac{dh}{dA_{ij}}$, $\frac{dh}{d\mathbf{x}^T}$, and such. In particular, we have the following nine cases:

	Scalar	**Vector**	**Matrix**
Scalar	$\frac{dy}{dx}$	$\frac{dy}{d\mathbf{x}}$	$\frac{dy}{d\mathbf{X}}$
Vector	$\frac{d\mathbf{y}}{dx}$	$\frac{d\mathbf{y}}{d\mathbf{x}}$	$\frac{d\mathbf{y}}{d\mathbf{X}}$
Matrix	$\frac{d\mathbf{Y}}{dx}$	$\frac{d\mathbf{Y}}{d\mathbf{x}}$	$\frac{d\mathbf{Y}}{d\mathbf{X}}$

We only define six of them; the derivatives of a scalar and a vector. Other cases are not required for 10-301/601. There are many different versions of definitions, but here we use the denominator-layout notation. Also note that we use d and ∂ interchangeably.

2.1.1 Derivatives of Scalar

We first consider when we take the derivative of a scalar.

1. *With respect to a scalar* (dy/dx): We already know this case. This is simply the single-variable function case.

2. *With respect to a vector* ($dy/d\mathbf{x}$): An example of this case is when $y = \|\mathbf{x}\|$. This is the gradient we defined. That is, for $\mathbf{x} \in \mathbb{R}^n$,

$$\frac{dy}{d\mathbf{x}} = \begin{bmatrix} dy/dx_1 \\ \vdots \\ dy/dx_n \end{bmatrix} \in \mathbb{R}^n = \mathbb{R}^{n \times 1}.$$

   We also define what happens when we take the derivative of a scalar with respect to a *row* vector $\mathbf{x}^T$:

$$\frac{dy}{d\mathbf{x}^T} = \begin{bmatrix} dy/dx_1 & \cdots & dy/dx_n \end{bmatrix} \in \mathbb{R}^{1 \times n}.$$

3. *With respect to a matrix* ($dy/d\mathbf{X}$): An example of this case is when $y = \sqrt{\sum_{i=1}^{m} \sum_{j=1}^{n} |X_{ij}|^2}$.*

*This is called the Frobenius norm, also denoted $\|\mathbf{X}\|_F$.

Expanding on the vector case, for $\mathbf{X} \in \mathbb{R}^{m \times n}$:

$$\frac{dy}{d\mathbf{X}} = \begin{bmatrix} dy/dX_{11} & \cdots & dy/dX_{1n} \\ \vdots & \ddots & \vdots \\ dy/dX_{m1} & \cdots & dy/dX_{mn} \end{bmatrix} \in \mathbb{R}^{m \times n}.$$

You will be asked to check if this is a valid generalization of the two definitions above as an exercise.

One thing to notice here is that when you take the derivative of a scalar, we end up with the same shape as the variable we took the derivative with respect to. For example, the shape of $dy/d\mathbf{x}$ is the same as the shape of $\mathbf{x}$. This is a nice property of the denominator-layout notation.

2.1.2 Derivatives of Vector

Now we expand the scalar case to vectors, i.e., $d\mathbf{y}/dx$, $d\mathbf{y}/d\mathbf{x}$, and $d\mathbf{y}/d\mathbf{X}$. Note that $\mathbf{y}$ here does not necessarily have to be a column vector. The exact same definitions apply to row vectors as well, including the resulting shapes.

1. *With respect to a scalar* ($d\mathbf{y}/dx$): An example of this case is $d(x\mathbf{v})/dx$ for a scalar x and constant vector $\mathbf{v} \in \mathbb{R}^n$. For $\mathbf{y} \in \mathbb{R}^n$, this is defined as:

$$\frac{d\mathbf{y}}{dx} = \begin{bmatrix} dy_1/dx & \cdots & dy_n/dx \end{bmatrix} \in \mathbb{R}^{1 \times n}.$$

2. *With respect to a vector* ($d\mathbf{y}/d\mathbf{x}$): An example of this case is $\mathbf{y} = \mathbf{A}\mathbf{x}$ for a constant matrix $\mathbf{A}$, and we evaluate $d\mathbf{y}/d\mathbf{x}$. For $\mathbf{y} \in \mathbb{R}^n$ and $\mathbf{x} \in \mathbb{R}^p$, this is defined as

$$\frac{d\mathbf{y}}{d\mathbf{x}} = \begin{bmatrix} \nabla y_1(x) & \nabla y_2(x) & \cdots & \nabla y_n(x) \end{bmatrix} = \begin{bmatrix} dy_1/dx_1 & dy_2/dx_1 & \cdots & dy_n/dx_1 \\ dy_1/dx_2 & dy_2/dx_2 & \cdots & dy_n/dx_2 \\ \vdots & \ddots & & \vdots \\ dy_1/dx_p & dy_2/dx_p & \cdots & dy_n/dx_p \end{bmatrix} \in \mathbb{R}^{p \times n}.$$

Consider when $\mathbf{y} = \mathbf{A}\mathbf{x}$ for a constant matrix $\mathbf{A} \in \mathbb{R}^{n \times p}$. Explicit multiplication yields

$$\begin{aligned} \mathbf{y} &= \mathbf{A}\mathbf{x} \\ &= \begin{bmatrix} A_{11} & \cdots & A_{1p} \\ \vdots & \ddots & \vdots \\ A_{n1} & \cdots & A_{np} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_p \end{bmatrix} \\ &= \begin{bmatrix} A_{11}x_1 + A_{12}x_2 + \cdots + A_{1p}x_p \\ \vdots \\ A_{n1}x_1 + A_{n2}x_2 + \cdots + A_{np}x_p \end{bmatrix} \\ &= \begin{bmatrix} \sum_{k=1}^{p} A_{1k}x_k \\ \vdots \\ \sum_{k=1}^{p} A_{nk}x_k \end{bmatrix}. \end{aligned}$$

This gives $y_i = \sum_{k=1}^{p} A_{ik}x_k$, and therefore $dy_i/dx_j = A_{ij}$. Hence, we have

$$\begin{aligned}\frac{d\mathbf{y}}{d\mathbf{x}} &= \begin{bmatrix} dy_1/dx_1 & dy_2/dx_1 & \cdots & dy_n/dx_1 \\ dy_1/dx_2 & dy_2/dx_2 & \cdots & dy_n/dx_2 \\ \vdots & & \ddots & \vdots \\ dy_1/dx_p & dy_2/dx_p & \cdots & dy_n/dx_p \end{bmatrix} \\ &= \begin{bmatrix} A_{11} & A_{21} & \cdots & A_{n1} \\ A_{12} & A_{22} & \cdots & A_{n2} \\ \vdots & & \ddots & \vdots \\ A_{1p} & A_{2p} & \cdots & A_{np} \end{bmatrix} \\ &= \mathbf{A}^T.\end{aligned}$$

Here we have derived one useful result:

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

3. *With respect to a matrix* ($d\mathbf{y}/d\mathbf{X}$): An example of this case is $\mathbf{y} = \mathbf{X}\mathbf{v}$ for a constant vector $\mathbf{v}$, and we evaluate $d\mathbf{y}/d\mathbf{X}$. In general, this encodes three dimensional information (dy_i/dX_{jk}) and is beyond the scope of this class. However, we define the following two specific cases that will be used throughout the class:

$$\frac{d\mathbf{X}\mathbf{v}}{d\mathbf{X}} = \mathbf{v}^T, \qquad \frac{d\mathbf{v}^T\mathbf{X}}{d\mathbf{X}} = \mathbf{v},$$

for a matrix $\mathbf{X}$ and constant vector $\mathbf{v}$. Note that the second case is the derivative of a row vector with respect to a matrix.

2.2 Chain Rule

Recall that for $h(x) = f(g(x))$ (single-variable functions), the chain rule was

$$\frac{dh}{dx} = \frac{df}{dg}\frac{dg}{dx} = \frac{dg}{dx}\frac{df}{dg}.$$

For the multivariable case $h(x) = f(g_1(x), g_2(x))$, the chain rule is extended as

$$\frac{dh}{dx} = \frac{\partial f}{\partial g_1}\frac{dg_1}{dx} + \frac{\partial f}{\partial g_2}\frac{dg_2}{dx} = \frac{dg_1}{dx}\frac{\partial f}{\partial g_1} + \frac{dg_2}{dx}\frac{\partial f}{\partial g_2}.$$

Visually, we can represent the two chain rules as Figure 2:

Figure 2: Chain rules visualized.

This can be thought of as adding all components that contribute to the change of h. Building on this, we can extend the chain rule to also work in matrix calculus.

Consider $\mathbf{x} \in \mathbb{R}^p$, $\mathbf{y} \in \mathbb{R}^r$, $\mathbf{z} \in \mathbb{R}^n$ where $\mathbf{z}$ is a function of $\mathbf{y}$, and $\mathbf{y}$ is a function of $\mathbf{x}$; that is, $\mathbf{z} = f(\mathbf{y})$, $\mathbf{y} = g(\mathbf{x})$, and therefore $\mathbf{z} = f(g(\mathbf{x}))$. We can visualize this as Figure 3. Note how this figure considers the most general possible case.

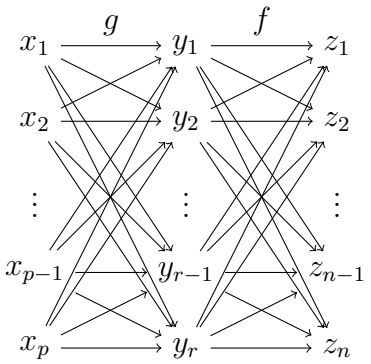


Figure 3: $\mathbf{z} = f(g(\mathbf{x}))$ visualized, where $\mathbf{z} = f(\mathbf{y})$ and $\mathbf{y} = g(\mathbf{x})$.

Now we derive the chain rule for vectors in matrix calculus. Recall that we have previously defined $d\mathbf{z}/d\mathbf{x}$ as

$$\frac{d\mathbf{z}}{d\mathbf{x}} = \begin{bmatrix} dz_1/dx_1 & dz_2/dx_1 & \cdots & dz_n/dx_1 \\ dz_1/dx_2 & dz_2/dx_2 & \cdots & dz_n/dx_2 \\ \vdots & & \ddots & \vdots \\ dz_1/dx_p & dz_2/dx_p & \cdots & dz_n/dx_p \end{bmatrix} \in \mathbb{R}^{p \times n}.$$

By the chain rule,

$$\frac{dz_i}{dx_j} = \sum_{k=1}^{r} \frac{dz_i}{dy_k} \frac{dy_k}{dx_j} = \sum_{k=1}^{r} \frac{dy_k}{dx_j} \frac{dz_i}{dy_k}.$$

This directly follows from Figure 4, which can be obtained by isolating only x_j and z_i from Figure 3:

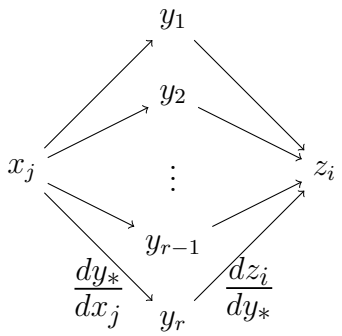


Figure 4: Chain rule visualized only considering z_i and x_j. y_* denotes any of $y_1, \cdots, y_r$.

Apply the scalar chain rule to each element of $d\mathbf{z}/d\mathbf{x}$. By the definition of matrix multiplication, observe that

$$\begin{aligned}
\left(\frac{d\mathbf{z}}{d\mathbf{x}}\right)^T &= \begin{bmatrix} dz_1/dx_1 & dz_1/dx_2 & \cdots & dz_1/dx_p \\ dz_2/dx_1 & dz_2/dx_2 & \cdots & dz_2/dx_p \\ \vdots & & \ddots & \vdots \\ dz_n/dx_1 & dz_n/dx_2 & \cdots & dz_n/dx_p \end{bmatrix} \in \mathbb{R}^{n\times p} \\
&= \begin{bmatrix} \sum_{k=1}^r \frac{dz_1}{dy_k}\frac{dy_k}{dx_1} & \sum_{k=1}^r \frac{dz_1}{dy_k}\frac{dy_k}{dx_2} & \cdots & \sum_{k=1}^r \frac{dz_1}{dy_k}\frac{dy_k}{dx_n} \\ \sum_{k=1}^r \frac{dz_2}{dy_k}\frac{dy_k}{dx_1} & \sum_{k=1}^r \frac{dz_2}{dy_k}\frac{dy_k}{dx_2} & \cdots & \sum_{k=1}^r \frac{dz_2}{dy_k}\frac{dy_k}{dx_n} \\ \vdots & & \ddots & \vdots \\ \sum_{k=1}^r \frac{dz_p}{dy_k}\frac{dy_k}{dx_1} & \sum_{k=1}^r \frac{dz_p}{dy_k}\frac{dy_k}{dx_2} & \cdots & \sum_{k=1}^r \frac{dz_p}{dy_k}\frac{dy_k}{dx_n} \end{bmatrix} \\
&= \begin{bmatrix} dz_1/dy_1 & dz_1/dy_2 & \cdots & dz_1/dy_r \\ dz_2/dy_1 & dz_2/dy_2 & \cdots & dz_2/dy_r \\ \vdots & & \ddots & \vdots \\ dz_n/dy_1 & dz_n/dy_2 & \cdots & dz_n/dy_r \end{bmatrix} \begin{bmatrix} dy_1/dx_1 & dy_1/dx_2 & \cdots & dy_1/dx_p \\ dy_2/dx_1 & dy_2/dx_2 & \cdots & dy_2/dx_p \\ \vdots & & \ddots & \vdots \\ dy_r/dx_1 & dy_r/dx_2 & \cdots & dy_r/dx_p \end{bmatrix} \\
&= \left(\frac{d\mathbf{z}}{d\mathbf{y}}\right)^T \left(\frac{d\mathbf{y}}{d\mathbf{x}}\right)^T.
\end{aligned}$$

Taking the transpose of both sides, we have that the chain rule extends to

$$\frac{d\mathbf{z}}{d\mathbf{x}} = \frac{d\mathbf{y}}{d\mathbf{x}}\frac{d\mathbf{z}}{d\mathbf{y}}.$$

Note the matrix multiplication order; $d\mathbf{y}/d\mathbf{x}$ comes first.* The order did not matter for the scalar case, but we need to be mindful of the order for the matrix case.

The key idea for this derivation was to manipulate the matrices cleverly and use the scalar chain rule. When other types of derivatives are involved, this chain rule may change; some derivatives may be transposed, and the multiplication order may change. The chain rules also vary depending on how the derivatives are defined. However, *the scalar chain rule must hold no matter what.*

*The chain rule is more natural using the numerator-layout notation, which is the transposed version of our notation (the chain rule is $d\mathbf{z}/d\mathbf{x} = (d\mathbf{z}/d\mathbf{y})(d\mathbf{y}/d\mathbf{x})$). This is one of the reasons why the transposed definitions are preferred by some.

2.3 Exercises

1. Recall the definition of the derivative of a scalar with respect to a matrix ($dy/d\mathbf{X}$). We will now check if this is a valid extension of the scalar and vector case. Evaluate the derivatives when $\mathbf{X} \in \mathbb{R}^{1\times 1}$, $\mathbf{X} \in \mathbb{R}^{n\times 1}$, and $\mathbf{X} \in \mathbb{R}^{1\times n}$. Which definition does each of them correspond to?

2. We have derived that $d(\mathbf{Ax})/d\mathbf{x} = \mathbf{A}^T$ for $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{A} \in \mathbb{R}^{n\times p}$ that does not depend on $\mathbf{x}$. $\mathbf{Ax}$ results in a vector, and thus we have used the $d\mathbf{y}/d\mathbf{x}$ definition. Now consider $d(\mathbf{x}^T\mathbf{B})/d\mathbf{x}$ for $\mathbf{B} \in \mathbb{R}^{p\times n}$. Recall that the definition of $d\mathbf{y}/d\mathbf{x}$ does not change even when $\mathbf{y}$ is a row vector. Evaluate $d(\mathbf{x}^T\mathbf{B})/d\mathbf{x}$.

3. The quadratic form $\mathbf{x}^T\mathbf{Ax}$ is a form we will encounter often.* In this question, we are interested in $d(\mathbf{x}^T\mathbf{Ax})/d\mathbf{x}$. Assume that $\mathbf{A}$ is not a function of $\mathbf{x}$.

   (a) Evaluate $\mathbf{x}^T\mathbf{Ax}$ when $\mathbf{x} = [x_1, x_2]^T$ and the (i, j)-th element of $\mathbf{A}$ is A_{ij}. Why do you think $\mathbf{x}^T\mathbf{Ax}$ is called the quadratic form?

   (b) Which definition of the derivative do we need in order to evaluate $d(\mathbf{x}^T\mathbf{Ax})/d\mathbf{x}$?

   (c) Assume $\mathbf{x} \in \mathbb{R}^2$ and $\mathbf{A} \in \mathbb{R}^{2\times 2}$. Evaluate $d(\mathbf{x}^T\mathbf{Ax})/d\mathbf{x}$.

   (d) Generalize the previous result to when $\mathbf{x} \in \mathbb{R}^n$ and $\mathbf{A} \in \mathbb{R}^{n\times n}$ and evaluate $d(\mathbf{x}^T\mathbf{Ax})/d\mathbf{x}$. Can you express the result in matrix form?

   (e) What happens when $\mathbf{A}$ is a symmetric matrix, i.e., $\mathbf{A}^T = \mathbf{A}$?

4. One of the most useful properties of differentiation is the linearity. That is, for scalar functions $f : \mathbb{R} \to \mathbb{R}$ and $g : \mathbb{R} \to \mathbb{R}$, we have $d(f(x)+g(x))/dx = df(x)/dx+dg(x)/dx$ and $d(a \cdot f(x))/dx = a \cdot df(x)/dx$ for some constant $a \in \mathbb{R}$. We will show that this extends to matrix calculus as well. Consider functions $u : \mathbb{R}^n \to \mathbb{R}^m$ and $v : \mathbb{R}^n \to \mathbb{R}^m$.

   (a) Show that for $\mathbf{x} \in \mathbb{R}^n$,

$$\frac{d(u(\mathbf{x}) + v(\mathbf{x}))}{d\mathbf{x}} = \frac{du(\mathbf{x})}{d\mathbf{x}} + \frac{dv(\mathbf{x})}{d\mathbf{x}}.$$

   (b) Show that for $\mathbf{x} \in \mathbb{R}^n$ and a constant $a \in \mathbb{R}$,

$$\frac{d(au(\mathbf{x}))}{d\mathbf{x}} = a\frac{du(\mathbf{x})}{d\mathbf{x}}.$$

5. Linear regression is the task of finding the "best" linear fit between labels $\mathbf{y} \in \mathbb{R}^n$ and attributes $\mathbf{X} \in \mathbb{R}^{m\times n}$. Concretely, we determine an adequate $\boldsymbol{\theta}$ such that $\mathbf{y} = \mathbf{X}\boldsymbol{\theta}$. One of the "best" choices of $\boldsymbol{\theta}$ is the one that minimizes the mean-squared error, which is given as

$$J(\boldsymbol{\theta}) = \frac{1}{N}(\mathbf{X}\boldsymbol{\theta} - \mathbf{y})^T(\mathbf{X}\boldsymbol{\theta} - \mathbf{y}).$$

   Find the $\boldsymbol{\theta}$ that minimizes the mean-squared error. As usual, this is the $\boldsymbol{\theta}$ such that $dJ(\boldsymbol{\theta})/d\boldsymbol{\theta} = 0$.

*Remember the definition (or one of the definitions) of positive-definite matrix?

6. Why can we write that

$$\frac{dh}{dx} = \frac{df}{dg}\frac{dg}{dx} = \frac{dg}{dx}\frac{df}{dg},$$

but not

$$\frac{d\mathbf{z}}{d\mathbf{x}} = \frac{d\mathbf{z}}{d\mathbf{y}}\frac{d\mathbf{y}}{d\mathbf{x}} = \frac{d\mathbf{y}}{d\mathbf{x}}\frac{d\mathbf{z}}{d\mathbf{y}}?$$

Which equality does not hold?

7. Evaluate $\partial f/\partial x$ and $\partial f/\partial y$ for each of the following:*

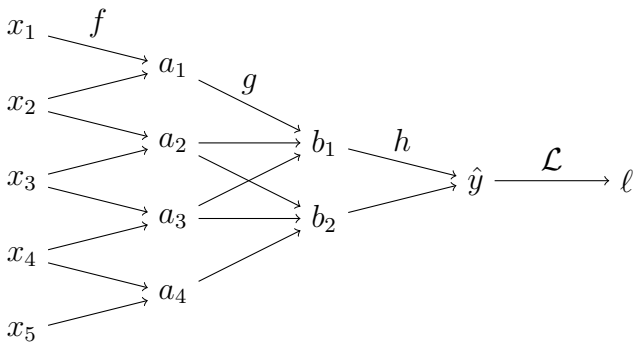
   (a) $f(u, v) = (u - v)e^u$, where $u = xy$ and $v = x^2 - y^2$

   (b) $f(u, v) = u \log v + v \log u$, where $u = \frac{x}{2} + \frac{2}{y}$ and $v = xe^y$

   (c) $f(u, v) = u \log v$, where $u = x \sin y + y \sin x$ and $v = x \cos y + y \cos x$

   (d) $f(u, v) = (u + v)/(1 - uv)$, where $u = \tan \frac{x+y}{2}$ and $v = \tan \frac{x-y}{2}$

8. Consider a neural network expressed as the following diagram:



This can be interpreted as a deep neural network with two hidden layers that accepts $\mathbf{x} \in \mathbb{R}^5$ as the input and outputs $\hat{y} \in \mathbb{R}$. The hidden layers $\mathbf{a}$ and $\mathbf{b}$ are computed as $\mathbf{a} = f(\mathbf{x})$ and $\mathbf{b} = g(\mathbf{a})$ for some functions $f : \mathbb{R}^5 \to \mathbb{R}^4$ and $g : \mathbb{R}^4 \to \mathbb{R}^2$. Finally, the output $\hat{y}$ is computed as $\hat{y} = h(\mathbf{b})$ for some function $h : \mathbb{R}^2 \to \mathbb{R}$. Now in order to update the network parameters, we perform gradient descent. The loss computed between the ground truth y and the current prediction output $\hat{y}$ is $\ell = \mathcal{L}(y, \hat{y})$.

   (a) Express $d\ell/db_1$ in terms of $dh/db_\square$ and $d\ell/d\hat{y}$. The blank ($\square$) must be filled in with either 1 or 2. Some terms may be reused with a different value in the blank.

   (b) Express $d\ell/d\mathbf{b}$ in terms of $dh/d\mathbf{b}$ and $d\ell/d\hat{y}$.

   (c) Express $d\ell/da_2$ in terms of $db_\square/da_\square$ and $d\ell/da_\square$. The blanks ($\square$) must be filled in with either 1 or 2. Some terms may be reused with a different value in the blank.

   (d) Express $d\ell/d\mathbf{a}$ in terms of $dg/d\mathbf{a}$ and $d\ell/d\mathbf{b}$.

*Questions taken almost directly from my undergraduate calculus book.

9. One of the extra readings for the neural network lecture is Deep Learning by Goodfellow, et al. The chain rule derivation in the book is as follows:

   > Suppose that $\mathbf{x} \in \mathbb{R}^m$, $\mathbf{y} \in \mathbb{R}^n$, g maps from $\mathbb{R}^m$ to $\mathbb{R}^n$, and f maps from $\mathbb{R}^n$ to $\mathbb{R}$. If $\mathbf{y} = g(\mathbf{x})$ and $z = f(\mathbf{y})$, then
   >
   > $$\frac{\partial z}{\partial x_i} = \sum_j \frac{\partial z}{\partial y_j} \frac{\partial y_j}{\partial x_i}. \tag{G6.45}$$
   >
   > In vector notation, this may be equivalently written as
   >
   > $$\nabla_{\mathbf{x}} z = \left(\frac{\partial \mathbf{y}}{\partial \mathbf{x}} \right)^T \nabla_{\mathbf{y}} z, \tag{G6.46}$$
   >
   > $\cdots$ *(omitted)*

   $\nabla_{\mathbf{x}} z$ and $\nabla_{\mathbf{y}} z$ are gradients of z with respect to $\mathbf{x}$ and $\mathbf{y}$, respectively. All vectors, including gradients, are column vectors.

   (a) Verify Eq. (G6.45) by applying the chain rule yourself.

   (b) Recall that the derivative convention we are using is called the denominator-layout notation. There is another set of definitions called the numerator-layout notation, which transposes all of the definitions we have. Is $\partial \mathbf{y}/\partial \mathbf{x}$ in Eq. (G6.46) defined using the denominator layout or the numerator layout? Explain why.

   (c) Derive Eq. (G6.46) from Eq. (G6.45) using the denominator layout and the numerator layout.

10. Hayden thinks it is odd that *all* definitions have to be transposed to build a different layout. Instead, he proposes a new set of definitions, which transposes *only* the derivative of a vector with respect to a vector ($d\mathbf{y}/d\mathbf{x}$).* He argues that this is more consistent with well-known mathematical concepts and therefore more convenient (this way, $dy/d\mathbf{x}$ is identical to the gradient and $d\mathbf{y}/d\mathbf{x}$ is identical to the Jacobian matrix).

    (a) For vectors $\mathbf{x}$, $\mathbf{y}$, $\mathbf{z}$ where $\mathbf{z}$ is a function of $\mathbf{y}$ and $\mathbf{y}$ is a function of $\mathbf{x}$, express $d\mathbf{z}/d\mathbf{x}$ in terms of $d\mathbf{z}/d\mathbf{y}$ and $d\mathbf{y}/d\mathbf{x}$ under this definition.

    (b) For vectors $\mathbf{x}$, $\mathbf{y}$ and scalar z where z is a function of $\mathbf{y}$ and $\mathbf{y}$ is a function of $\mathbf{x}$, express $dz/d\mathbf{x}$ in terms of $dz/d\mathbf{y}$ and $d\mathbf{y}/d\mathbf{x}$ under this definition.

    (c) Compare the result of (b) with Eq. (G6.46) in Question 9. Describe one caveat of this definition.

*Some authors actually use this.

3 Computing the Derivatives

In this section, we focus on how to actually compute various derivatives. We will first cover the "hacky" way which usually suffices for 10-301/601, and the mathematically rigorous way in case the hacky method fails.

3.1 Shape Matching

One thing we can take advantage of matrix multiplication is that it is defined only when the shapes of the operands match. Recall that for two matrices $\mathbf{X} \in \mathbb{R}^{m \times n}$ and $\mathbf{Y} \in \mathbb{R}^{n \times p}$, $\mathbf{Z} = \mathbf{XY} \in \mathbb{R}^{m \times p}$ is defined as

$$\mathbf{Z} = (Z_{ij}), \text{ where } Z_{ij} = \sum_{k=1}^{n} X_{ik} Y_{kj}.$$

Note the shapes of $\mathbf{X}$ and $\mathbf{Y}$. The number of columns of $\mathbf{X}$ and the number of rows of $\mathbf{Y}$ have to be equal for $\mathbf{XY}$ to be defined. The resultant product has the same number of rows as $\mathbf{X}$ and the same number of column as $\mathbf{Y}$.

With this and the scalar version of the chain rule, we can "derive" the vector chain rule. Consider $\mathbf{x} \in \mathbb{R}^p$, $\mathbf{y} \in \mathbb{R}^r$, $\mathbf{z} \in \mathbb{R}^n$ where $\mathbf{z}$ is a function of $\mathbf{y}$, and $\mathbf{y}$ is a function of $\mathbf{x}$, and we derive $d\mathbf{z}/d\mathbf{x}$ again in this setting. If $\mathbf{x}$, $\mathbf{y}$, and $\mathbf{z}$ were all scalars, dz/dx simply would be

$$\frac{dz}{dx} = \frac{dz}{dy}\frac{dy}{dx}.$$

From here, we can guess that $d\mathbf{z}/d\mathbf{x}$ would be a product of $d\mathbf{z}/d\mathbf{y}$ and $d\mathbf{y}/d\mathbf{x}$. We also know that the shapes of $d\mathbf{z}/d\mathbf{x}$, $d\mathbf{z}/d\mathbf{y}$, and $d\mathbf{y}/d\mathbf{x}$ are $p \times n$, $r \times n$, and $p \times r$, respectively. Therefore, the correct order of multiplication is

$$\frac{d\mathbf{z}}{d\mathbf{x}} = \frac{d\mathbf{y}}{d\mathbf{x}}\frac{d\mathbf{z}}{d\mathbf{y}}.$$

The new chain rule "derivation" is not rigorous, and technically is not even a proper proof. However, this shaping matching technique is extremely useful for sanity check (and maybe also multiple-choice questions; sometimes you can eliminate some options with incorrect shapes). Typically, the general procedure for this would be:

1. Determine what to evaluate. You may have to do this yourself, or the question may tell you explicitly.

2. Identify the shape of the final answer. If you are taking the derivative of a scalar, the shape is the same as the shape of the variable you are taking the derivative with respect to. If you are taking the derivative of an n-dimensional vector, the shape is something by n.

3. For multiple choice questions, eliminate any options whose shape does not match or the operation is not defined. This includes those multiplying or adding matrices of wrong shapes.

4. If you can exactly determine what terms and factors you need, you may be able to obtain the answer by transposing and matching them until all operations are properly defined and the final shape is correct.

Of course, this is closer to guessing the answer rather than logically deriving it. Also, this may fail if the shapes *happen to* match. For example, for $\mathbf{x}, \mathbf{y}, \mathbf{z} \in \mathbb{R}^n$, $d\mathbf{z}/d\mathbf{x}$, $d\mathbf{z}/d\mathbf{y}$, $d\mathbf{y}/d\mathbf{x}$ are all $n \times n$. Selecting and multiplying any two of them in any order is still valid as the shapes are fine, but the answer will be incorrect. Also, this method cannot be used for any operations that do not change the shape, such as addition, subtraction, and scalar multiplication.

3.2 Generalizing Single Element

A more logically correct and mathematically rigorous way is to consider a single element of a matrix, and generalize it to obtain the full matrix. Consider the following four cases, which were the only non-scalar derivative definitions we have:

1. Case $dy/d\mathbf{x}$ (or $dy/d\mathbf{x}^T$): the i-th element is dy/dx_i.
2. Case $dy/d\mathbf{X}$: the (i,j)-th element is dy/dX_{ij}.
3. Case $d\mathbf{y}/dx$: the i-th element is dy_i/dx.
4. Case $d\mathbf{y}/d\mathbf{x}$: the (i,j)-th element is dy_j/dx_i (*not* dy_i/dx_j).

As an example, we will derive $d(\mathbf{Ax})/d\mathbf{x} = \mathbf{A}^T$ again here for $\mathbf{x} \in \mathbb{R}^p$ and some constant matrix $\mathbf{A} \in \mathbb{R}^{n \times p}$. Let $\mathbf{y} = \mathbf{Ax} \in \mathbb{R}^n$ for convenience. Earlier we obtained this by explicitly computing everything. Here we will try and simplify this by considering only one entry of $d\mathbf{y}/d\mathbf{x}$.

Say we compute one of the elements of $d\mathbf{y}/d\mathbf{x}$ first; the (i,j)-th one, or dy_j/dx_i. Through this, we have reduced the problem to simple scalar differentiation. Now we need to identify what y_j is. By the definition of matrix multiplication,

$$\begin{aligned} y_j &= y_{j1} \\ &= \sum_{k=1}^{p} A_{jk} x_{k1} \\ &= \sum_{k=1}^{p} A_{jk} x_k. \end{aligned}$$

Here we interpreted $\mathbf{x}$ and $\mathbf{y}$ as (vector dimension) $\times$ 1 matrices as necessary. Then we have

$$\begin{aligned} \frac{dy_j}{dx_i} &= \frac{d}{dx_i} \sum_{k=1}^{p} A_{jk} x_k \\ &= A_{ji}. \end{aligned}$$

This is the (i,j)-th element of the desired derivative. The matrix whose (i,j)-th element is A_{ji} is $\mathbf{A}^T$, so we conclude that

$$\frac{d\mathbf{y}}{d\mathbf{x}} = \mathbf{A}^T.$$

This method is clearly logically sound and mathematically solid. Another advantage of this method is that this works for any definition of matrix derivatives as long as we change the indices accordingly. However, this is more difficult than simple shape matching, and thinking in terms of indices and one element in a matrix can be tricky.

Also, this can be extended to derivatives of any dimensions. For example, consider we take the derivative of a 5D tensor $\mathbf{T}$ with respect to a matrix $\mathbf{X}$. There are a total of 7 dimensions where values can change, so one element of the "naïve" derivative would be dT_{ijklm}/dX_{xy}. However, not all seven dimensions are necessarily required (i.e., fewer *free variables* may suffice). Some elements may always have the same value (usually zero), and some rows/columns/elements may be repeated. We may choose to omit these pieces of redundant information as you will see in the exercises.

We have briefly mentioned that the shape matching method fails when the operations applied do not change the shape. It is easy to see that the single element method can be used instead. In fact, we can utilize it for any arbitrary well-defined operations. One extremely common and handy operation in machine learning is element-wise multiplication, also called the Hadamard product, denoted $\odot$. This is also detailed in one of the exercise questions (***do not skip this question***).

3.3 Matrix Multiplication Review

Matrix multiplication is simple, but we rarely think about the index-based definition. However, it is crucial to read sums and/or products of scalars and translate them back to matrix operations in order to use the single element method. To this end, we will review how matrix multiplication and some common more specific cases are defined.

1. For $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, $d = \mathbf{a}^T\mathbf{b} \in \mathbb{R}$ is defined as:

$$d = \sum_{k=1}^{n} a_k b_k.$$

2. For $\mathbf{a}, \mathbf{b} \in \mathbb{R}^n$, $\mathbf{D} = \mathbf{a}\mathbf{b}^T \in \mathbb{R}^{n\times n}$ is defined as:

$$D_{ij} = a_i b_j.$$

3. For $\mathbf{A} \in \mathbb{R}^{m\times p}$ and $\mathbf{v} \in \mathbb{R}^p$, $\mathbf{w} = \mathbf{A}\mathbf{v} \in \mathbb{R}^m$ is defined as

$$w_i = \sum_{k=1}^{p} A_{ik} v_k.$$

This can be visualized as the following:

$$\mathbf{A}\mathbf{v} = \begin{bmatrix} — & \mathbf{A}_{1,:}{}^T & — \\ & \vdots & \\ — & \mathbf{A}_{m,:}{}^T & — \end{bmatrix} \mathbf{v} = \begin{bmatrix} \mathbf{A}_{1,:}{}^T\mathbf{v} \\ \vdots \\ \mathbf{A}_{m,:}{}^T\mathbf{v} \end{bmatrix},$$

which gives rise to

$$w_i = \mathbf{A}_{i,:}{}^T \mathbf{v}.$$

Notice the transpose operator. The usual convention is to write $\mathbf{A}_{i,:}$ as a *column* vector even though it is the i-th *row* of $\mathbf{A}$. Here we use the transpose operator (for this section) to explicitly state that the row selected is represented as a row vector.

The following visualization is also possible:

$$\mathbf{A}\mathbf{v} = \begin{bmatrix} | & & | \\ \mathbf{A}_{:,1} & \cdots & \mathbf{A}_{:,p} \\ | & & | \end{bmatrix} \begin{bmatrix} v_1 \\ \vdots \\ v_p \end{bmatrix} = \begin{bmatrix} | \\ \mathbf{A}_{:,1} \\ | \end{bmatrix} v_1 + \cdots + \begin{bmatrix} | \\ \mathbf{A}_{:,p} \\ | \end{bmatrix} v_p,$$

which yields

$$\mathbf{A}\mathbf{v} = \sum_{k=1}^{p} \mathbf{A}_{:,k} v_k = \sum_{k=1}^{p} v_k \mathbf{A}_{:,k}.$$

Note that v_k can be multiplied both before and after $\mathbf{A}_{:,k}$ only because v_k is a scalar.

4. For $\mathbf{A} \in \mathbb{R}^{m \times p}$ and $\mathbf{u} \in \mathbb{R}^m$, $\mathbf{y} = \mathbf{u}^T \mathbf{A} \in \mathbb{R}^{1 \times p}$ is defined as

$$y_i = y_{1i} = \sum_{k=1}^{m} u_k A_{ki}.$$

We try similar visualizations. Considering each column of $\mathbf{A}$ gives

$$\mathbf{u}^T \mathbf{A} = \mathbf{u}^T \begin{bmatrix} | & & | \\ \mathbf{A}_{:,1} & \cdots & \mathbf{A}_{:,p} \\ | & & | \end{bmatrix} = \begin{bmatrix} \mathbf{u}^T \mathbf{A}_{:,1} & \cdots & \mathbf{u}^T \mathbf{A}_{:,m} \end{bmatrix},$$

which can be interpreted as

$$y_i = y_{1i} = \mathbf{u}^T \mathbf{A}_{:,i}.$$

Similarly, focusing on the rows of $\mathbf{A}$, we have

$$\begin{aligned} \mathbf{u}^T \mathbf{A} &= \begin{bmatrix} u_1 & \cdots & u_m \end{bmatrix} \begin{bmatrix} — & \mathbf{A}_{1,:}{}^T & — \\ & \vdots & \\ — & \mathbf{A}_{m,:}{}^T & — \end{bmatrix} \\ &= u_1 \begin{bmatrix} — & \mathbf{A}_{1,:}{}^T & — \end{bmatrix} + \cdots + u_m \begin{bmatrix} — & \mathbf{A}_{m,:}{}^T & — \end{bmatrix}, \end{aligned}$$

which is equivalent to

$$\mathbf{u}^T \mathbf{A} = \sum_{k=1}^{m} u_k \mathbf{A}_{k,:} = \sum_{k=1}^{m} \mathbf{A}_{k,:} u_k.$$

Again, u_k can come both before and after $\mathbf{A}_{k,:}$ only because u_k is a scalar.

5. For $\mathbf{A} \in \mathbb{R}^{m \times p}$ and $\mathbf{B} \in \mathbb{R}^{p \times n}$, $\mathbf{C} = \mathbf{AB} \in \mathbb{R}^{m \times n}$ is defined as

$$C_{ij} = \sum_{k=1}^{p} A_{ik} B_{kj}.$$

Again, this can be thought of as

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} — & \mathbf{A}_{1,:}{}^T & — \\ & \vdots & \\ — & \mathbf{A}_{m,:}{}^T & — \end{bmatrix} \begin{bmatrix} | & & | \\ \mathbf{B}_{:,1} & \cdots & \mathbf{B}_{:,n} \\ | & & | \end{bmatrix} = \begin{bmatrix} \mathbf{A}_{1,:}{}^T\mathbf{B}_{:,1} & \cdots & \mathbf{A}_{1,:}{}^T\mathbf{B}_{:,n} \\ \vdots & \ddots & \vdots \\ \mathbf{A}_{m,:}{}^T\mathbf{B}_{:,1} & \cdots & \mathbf{A}_{m,:}{}^T\mathbf{B}_{:,n} \end{bmatrix},$$

and we derive the expression

$$C_{ij} = \mathbf{A}_{i,:}{}^T\mathbf{B}_{:,j},$$

or from a different perspective,

$$\begin{aligned} \mathbf{C} = \mathbf{AB} &= \begin{bmatrix} | & & | \\ \mathbf{A}_{:,1} & \cdots & \mathbf{A}_{:,p} \\ | & & | \end{bmatrix} \begin{bmatrix} — & \mathbf{B}_{1,:}{}^T & — \\ & \vdots & \\ — & \mathbf{B}_{p,:}{}^T & — \end{bmatrix} \\ &= \begin{bmatrix} | \\ \mathbf{A}_{:,1} \\ | \end{bmatrix} \begin{bmatrix} — & \mathbf{B}_{1,:}{}^T & — \end{bmatrix} + \cdots + \begin{bmatrix} | \\ \mathbf{A}_{:,p} \\ | \end{bmatrix} \begin{bmatrix} — & \mathbf{B}_{p,:}{}^T & — \end{bmatrix}, \end{aligned}$$

which gives

$$\mathbf{C} = \sum_{k=1}^{p} \mathbf{A}_{:,p}\mathbf{B}_{p,:}{}^T.$$

These visualizations break apart both matrices into vectors. Now we try leaving one of the matrices as is, which yields the following:

$$\mathbf{C} = \mathbf{AB} = \mathbf{A} \begin{bmatrix} | & & | \\ \mathbf{B}_{:,1} & \cdots & \mathbf{B}_{:,n} \\ | & & | \end{bmatrix} = \begin{bmatrix} | & & | \\ \mathbf{AB}_{:,1} & \cdots & \mathbf{AB}_{:,n} \\ | & & | \end{bmatrix},$$

in other words,

$$\mathbf{C}_{:,i} = \mathbf{AB}_{:,i}.$$

Similarly, leaving $\mathbf{B}$ as is,

$$\mathbf{C} = \mathbf{AB} = \begin{bmatrix} — & \mathbf{A}_{1,:}{}^T & — \\ & \vdots & \\ — & \mathbf{A}_{m,:}{}^T & — \end{bmatrix} \mathbf{B} = \begin{bmatrix} — & \mathbf{A}_{1,:}{}^T\mathbf{B} & — \\ & \vdots & \\ — & \mathbf{A}_{m,:}{}^T\mathbf{B} & — \end{bmatrix},$$

and we have the final interpretation:

$$\mathbf{C}_{i,:}{}^T = \mathbf{A}_{i,:}{}^T\mathbf{B}.$$

Finally, remember that matrix multiplication is not commutative, but associative. It is extremely easy to show that it is not commutative; pick any two arbitrary matrices and likely they will work as a counterexample. Associativity can be shown by comparing the (i, j)-th element of $(\mathbf{AB})\mathbf{C}$ and $\mathbf{A}(\mathbf{BC})$.

3.4 Exercises

1. Recall for scalar functions f and g, the product rule is $(fg)' = f'g + fg'$. In this problem, we extend this to vector functions. Consider vectors $\mathbf{u}, \mathbf{v} \in \mathbb{R}^n$ that are functions of $\mathbf{x} \in \mathbb{R}^m$.* We are interested in $d(\mathbf{u}^T\mathbf{v})/d\mathbf{x}$.

   (a) What is the expected shape of $d(\mathbf{u}^T\mathbf{v})/d\mathbf{x}$?

   (b) **Select all that apply:** Which of the following have the same shape as $d(\mathbf{u}^T\mathbf{v})/d\mathbf{x}$?

   ☐ $\frac{d\mathbf{u}}{d\mathbf{x}}\frac{d\mathbf{v}}{d\mathbf{x}}$

   ☐ $\frac{d\mathbf{v}}{d\mathbf{x}}\frac{d\mathbf{u}}{d\mathbf{x}}$

   ☐ $\mathbf{u}\frac{d\mathbf{u}}{d\mathbf{x}} + \mathbf{v}\frac{d\mathbf{v}}{d\mathbf{x}}$

   ☐ $\mathbf{v}\frac{d\mathbf{u}}{d\mathbf{x}} + \mathbf{u}\frac{d\mathbf{v}}{d\mathbf{x}}$

   ☐ $\frac{d\mathbf{u}}{d\mathbf{x}}\mathbf{v} + \mathbf{u}\frac{d\mathbf{v}}{d\mathbf{x}}$

   ☐ $\frac{d\mathbf{u}}{d\mathbf{x}}\mathbf{v} - \mathbf{u}\frac{d\mathbf{v}}{d\mathbf{x}}$

   ☐ $\frac{d\mathbf{u}}{d\mathbf{x}}\mathbf{v} + \frac{d\mathbf{v}}{d\mathbf{x}}\mathbf{u}$

   ☐ $\frac{d\mathbf{u}}{d\mathbf{x}}\mathbf{u} + \frac{d\mathbf{v}}{d\mathbf{x}}\mathbf{v}$

   (c) What is one element of $d(\mathbf{u}^T\mathbf{v})/d\mathbf{x}$? Also specify the index of that element.

   (d) Generalize the answer of (c) to evaluate $d(\mathbf{u}^T\mathbf{v})/d\mathbf{x}$.

2. Element-wise operations are very common for vectors and matrices. We will explore the derivatives when these operations are involved.

   (a) For $\mathbf{x} \in \mathbb{R}^n$ and $f : \mathbb{R} \to \mathbb{R}$, define $\mathbf{y}$ as $y_i = f(x_i)$. Evaluate $d\mathbf{y}/d\mathbf{x}$.

   (b) For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, define $\mathbf{z} = \mathbf{x} \odot \mathbf{y}$ as $z_i = x_i y_i$. Evaluate $d\mathbf{z}/d\mathbf{x}$.

   (c) For $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$, define $\mathbf{z}$ as $z_i = f(x_i, y_i)$. Evaluate $d\mathbf{z}/d\mathbf{x}$.

3. We have previously defined that $d(\mathbf{X}\mathbf{v})/d\mathbf{X} = \mathbf{v}^T$ without reasoning about it; we will justify this definition in this problem. Consider a matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$ and a constant vector $\mathbf{v} \in \mathbb{R}^n$. $\mathbf{y} \in \mathbb{R}^m$ is defined as $\mathbf{y} = \mathbf{X}\mathbf{v}$.

   (a) Observe that some elements of $d(\mathbf{X}\mathbf{v})/d\mathbf{X}$ are always zero. What is the relation of i, j and k when dy_i/dX_{jk} is necessarily zero?

   (b) Evaluate dy_i/dX_{jk} only where it can be nonzero; i.e., evaluate the value only where i, j, and k do *not* satisfy (a).

   (c) Say we only want to compute $d\mathbf{y}/d\mathbf{X}$ for only where dy_i/dX_{jk} can be nonzero. Argue that this information can be represented as a 2D matrix.

   (d) Argue further that the matrix in (c) can be represented as a vector.

   (e) Construct the vector in (d) so that $d\mathbf{y}/d\mathbf{X} = d(\mathbf{X}\mathbf{v})/d\mathbf{X} = \mathbf{v}^T$.

   (f) After all, we forced the vector in (d) to fit our definition. We might as well have transposed the matrix in (c) and said $d(\mathbf{X}\mathbf{v})/d\mathbf{X} = \mathbf{v}$. What is one advantage of *not* doing so; i.e., why define $d(\mathbf{X}\mathbf{v})/d\mathbf{X} = \mathbf{v}^T$?

*You can consider this as $\mathbf{u} = f(\mathbf{x})$, $\mathbf{v} = g(\mathbf{x})$ for some functions $f, g : \mathbb{R}^m \to \mathbb{R}^n$.

4. The chain rule can be extended to when derivatives of a scalar with respect to a matrix are involved. We will specifically consider the case where $z \in \mathbb{R}$ is a function of row vector $\mathbf{y}$, where $\mathbf{y} = \mathbf{v}^T\mathbf{X}$ for some matrix $\mathbf{X} \in \mathbb{R}^{m\times n}$ and a constant vector $\mathbf{v} \in \mathbb{R}^m$. Then by the chain rule, $dz/d\mathbf{X}$ can be represented as a product of $d\mathbf{y}/d\mathbf{X}$ and $dz/d\mathbf{y}$.

   (a) Derive the chain rule by matching the shapes. *Hint:* $d\mathbf{y}/d\mathbf{X} = d(\mathbf{v}^T\mathbf{X})/d\mathbf{X} =$?

   (b) Evaluate $dz/d\mathbf{X}$ by considering one element of $dz/d\mathbf{X}$ first, then use the scalar chain rule. Do *not* use the known result of $d\mathbf{y}/d\mathbf{X}$ itself.

   (c) Compare the answers to (a) and (b) and justify our definition of $d(\mathbf{v}^T\mathbf{X})/d\mathbf{X}$.

5. Consider a neural network that accepts $\mathbf{x} \in \mathbb{R}^n$ as the input and outputs $\hat{\mathbf{y}} \in \mathbb{R}^m$. The intermediate activations $\mathbf{a}$, the output $\hat{\mathbf{y}}$, and the loss $\ell \in \mathbb{R}$ between $\hat{\mathbf{y}}$ and the ground truth $\mathbf{y} \in \mathbb{R}^m$ are calculated as follows:

$$\begin{aligned}\mathbf{a} &= f(\mathbf{M}\mathbf{x}),\\ \hat{\mathbf{y}} &= g(\mathbf{N}\mathbf{a}),\\ \ell &= h(\hat{\mathbf{y}}, \mathbf{y}),\end{aligned}$$

   for matrices $\mathbf{M} \in \mathbb{R}^{p\times n}$ and $\mathbf{N} \in \mathbb{R}^{m\times p}$, and functions $f, g : \mathbb{R} \to \mathbb{R}$ and $h : \mathbb{R}^m \to \mathbb{R}$. f and g are applied element-wise.

   (a) Express $d\ell/d\mathbf{N}$ in terms of $d\ell/d\hat{\mathbf{y}}$ and $d\hat{\mathbf{y}}/d\mathbf{N}$.

   (b) Express your answer in (a) with the only derivative not evaluated being $d\ell/d\hat{\mathbf{y}}$.

   (c) Fill in the blanks ($\square$) so that the equality holds:

$$\frac{d\ell}{d\mathbf{M}} = \frac{d\square}{d\square}\frac{d\square}{d\square}\frac{d\square}{d\square}.$$

   The blanks can only be one of ℓ, $\mathbf{a}$, $\mathbf{x}$, $\mathbf{y}$, $\hat{\mathbf{y}}$, $\mathbf{M}$, and $\mathbf{N}$. Some may be reused. One of the factors must be $d\ell/d\hat{\mathbf{y}}$.

   (d) Express your answer in (c) with the only derivative not evaluated being $d\ell/d\hat{\mathbf{y}}$.

   (e) In practice, neural networks are often updated by using a set of inputs $\mathbf{x}_1, \cdots, \mathbf{x}_B$.* The final loss used for the network update is the arithmetic mean of individual losses obtained by passing the inputs one by one. Describe how $d\ell/d\mathbf{M}$ and $d\ell/d\mathbf{N}$ change under this setting.

*This set of inputs is called a batch or a mini-batch, and the subscript B is from "Batch size."

6. The forward-backward algorithm for HMM (Hidden Markov Model) follows the update rules given as

$$\alpha_t(j) = A_{jx_t} \sum_{k=1}^{J} \alpha_{t-1}(k) B_{kj},$$
$$\beta_t(j) = \sum_{k=1}^{J} A_{kx_{t+1}} \beta_{t+1}(k) B_{jk}.$$

Conceptual understanding of the algorithm is not required for this question. Interpret $\mathbf{A}$ as a $J \times W$ matrix, and $\mathbf{B}$ as a $J \times J$ matrix. x_t is a fixed sequence of integers in range $[1, W]$. j is an integer in range $[1, J]$. J and W are fixed integers. Assume that α_t, α_{t-1}, β_t, and β_{t+1} are all well-defined (i.e., ignore cases such as when $t = 0$).

When we use these rules as they are, we have to iterate over every single possible j and update the values one by one. In this problem, we are interested in deriving the matrix form of these rules so that we can update them all at the same time. Concretely, we define vectors $\boldsymbol{\alpha}_t$ and $\boldsymbol{\beta}_t$ as

$$\boldsymbol{\alpha}_t = [\alpha_t(1), \cdots, \alpha_t(J)]^T,$$
$$\boldsymbol{\beta}_t = [\beta_t(1), \cdots, \beta_t(J)]^T$$

and you are to derive new update rules

$$\boldsymbol{\alpha}_t = (\text{some expression involving } \boldsymbol{\alpha}_{t-1}),$$
$$\boldsymbol{\beta}_t = (\text{some expression involving } \boldsymbol{\beta}_{t+1})$$

which are equivalent to the original update rules. Express the update rules in matrix form.

7. Consider column vectors $\mathbf{x}, \mathbf{y} \in \mathbb{R}^n$ and matrices $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{n \times n}$. Each of the following is one element, row, or vector of a vector or matrix obtained by multiplying some of $\mathbf{x}$, $\mathbf{y}$, $\mathbf{A}$, $\mathbf{B}$, $\mathbf{x}^T$, $\mathbf{y}^T$, $\mathbf{A}^T$, and $\mathbf{B}^T$, or the result itself. Write the matrix multiplication forms generalizing the following expressions. Any vectors given in the following are column vectors. For example, $\mathbf{A}_{i,:}$ is the i-th row of $\mathbf{A}$ as a column vector.

(a) $\sum_{i=1}^{N} A_{ij} x_i y_i$

(b) $\sum_{i=1}^{N} A_{ij} x_i$

(c) $\sum_{j=1}^{N} A_{ij} x_j y_j$

(d) $\sum_{j=1}^{N} \sum_{k=1}^{N} A_{ki} B_{kj} x_j$

(e) $\sum_{i=1}^{N} \mathbf{x}^T \mathbf{A}_{i,:} \mathbf{A}_{i,:}^T$

(f) $\sum_{j=1}^{N} A_{ij} x_i y_j$

(g) $\sum_{i=1}^{N} \mathbf{A}_{:,i} \mathbf{B}_{:,i}^T \mathbf{x}$

(h) $x_i y_j A_{ij}$

4.2 Section 2

1. Same as dy/dx, $dy/d\mathbf{x}$, and $dy/d\mathbf{x}^T$, respectively. Note that $\mathbb{R}^{1\times 1}$ is a scalar, $\mathbb{R}^{n\times 1}$ is a column vector, and $\mathbb{R}^{1\times n}$ is a row vector, so this result is expected.

2. Similar to how we derived the $d(\mathbf{Ax})/d\mathbf{x}$ case, we directly compute $\mathbf{x}^T\mathbf{B}$ first.

$$\begin{aligned}\mathbf{x}^T\mathbf{B} &= \begin{bmatrix} x_1 & \cdots & x_p \end{bmatrix} \begin{bmatrix} B_{11} & \cdots & B_{1n} \\ \vdots & \ddots & \vdots \\ B_{p1} & \cdots & B_{pn} \end{bmatrix} \\ &= \begin{bmatrix} B_{11}x_1 + B_{21}x_2 + \cdots + B_{p1}x_p & \cdots & B_{1n}x_1 + B_{2n}x_2 + \cdots + B_{pn}x_p \end{bmatrix} \\ &= \begin{bmatrix} \sum_{k=1}^{p} B_{k1}x_k & \cdots & \sum_{k=1}^{p} B_{kn}x_k \end{bmatrix}.\end{aligned}$$

   The i-th element of $\mathbf{x}^T\mathbf{B}$ is $\sum_{k=1}^{p} B_{ki}x_k$, and therefore $dy_i/dx_j = B_{ji}$ ($\mathbf{y} = \mathbf{x}^T\mathbf{B}$). Hence, we have

$$\begin{aligned}\frac{d(\mathbf{x}^T\mathbf{B})}{d\mathbf{x}} &= \begin{bmatrix} dy_1/dx_1 & dy_2/dx_1 & \cdots & dy_n/dx_1 \\ dy_1/dx_2 & dy_2/dx_2 & \cdots & dy_n/dx_2 \\ \vdots & & \ddots & \vdots \\ dy_1/dx_p & dy_2/dx_p & \cdots & dy_n/dx_p \end{bmatrix} \\ &= \begin{bmatrix} B_{11} & B_{12} & \cdots & B_{1n} \\ B_{21} & B_{22} & \cdots & B_{2n} \\ \vdots & & \ddots & \vdots \\ B_{p1} & B_{p2} & \cdots & B_{pn} \end{bmatrix} \\ &= \mathbf{B}.\end{aligned}$$

3. (a)

$$\begin{aligned}\mathbf{x}^T\mathbf{Ax} &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\ &= \begin{bmatrix} x_1 & x_2 \end{bmatrix} \begin{bmatrix} A_{11}x_1 + A_{12}x_2 \\ A_{21}x_1 + A_{22}x_2 \end{bmatrix} \\ &= A_{11}x_1^2 + (A_{12} + A_{21})x_1x_2 + A_{22}x_2^2\end{aligned}$$

   Each term is a polynomial of degree 2, so this is a quadratic.

   (b) $\mathbf{x}^T\mathbf{Ax}$ is a scalar and $\mathbf{x}$ is a vector. Therefore, we need the definition of the derivative of a scalar with respect to a vector.

   (c) Under this assumption, $\mathbf{x}^T\mathbf{Ax} = A_{11}x_1^2 + (A_{12} + A_{21})x_1x_2 + A_{22}x_2^2$ as we found in

End of main content

Solutions on the next page

4 Solutions

Clicking a question number will take you to that question.

4.1 Section 1

1. (a) $\sigma'(x) = ((1+e^{-x})^{-1})' = e^{-x}/(1+e^{-x})^2$

   (b) $e^{-x}/(1+e^{-x})^2 = (1+e^{-x}-1)/(1+e^{-x})^2 = \sigma(x)(1-\sigma(x))$

   (c) $\tanh'(x) = ((e^x+e^{-x})^2 - (e^x-e^{-x})^2)/(e^x+e^{-x})^2$

   (d) $((e^x+e^{-x})^2 - (e^x-e^{-x})^2)/(e^x+e^{-x})^2 = 1-\tanh^2(x)$

   Please remember the results of (b) and (d), just like how you can say directly from memory that $\sin'(x) = \cos(x)$. $\sigma'(x) = \sigma(x)(1-\sigma(x))$ and $\tanh'(x) = 1-\tanh^2(x)$ without derivation from now on.

2. (a) $yx^{y-1} + y^x \log y$, $x^y \log x + xy^{x-1}$

   (b) $\sin(x)(-\cos(y+\cos x))$, $\cos(y+\cos x)$

   (c) $y(e^{xy}+1/x)$, $xe^{xy} + \log 3x$

   (d) $-y^2[\sin(xy)+\cos(xy)]$, $-xy\sin(xy) - \sin(xy) - xy\cos(xy) + \cos(xy)$, $-xy\sin(xy) - \sin(xy) - xy\cos(xy) + \cos(xy)$, $-x^2[\sin(xy)+\cos(xy)]$

   (e) $\log(y)\cdot x^{-1+\log y}$, $2 + \log(x)x^{\log(y)}(1/y)$

   (f) $2(x+y)$, $2(x+y)$

   (g) $2x_i$

   (h) w_i

   The most important ones here are (g) and (h).

3. (a) $[y(2x+y), x(x+2y)]^T$

   (b) $[2(x+y), 2(x+y)]^T$

   (c) $\begin{bmatrix} y^2e^{xy}\left[\cos(e^{xy}) - e^{xy}\sin(e^{xy})\right] & e^{xy}\left[\cos(e^{xy})(1+xy) - xye^{xy}\sin(e^{xy})\right] \\ e^{xy}\left[\cos(e^{xy})(1+xy) - xye^{xy}\sin(e^{xy})\right] & x^2e^{xy}\left[\cos(e^{xy}) - e^{xy}\sin(e^{xy})\right] \end{bmatrix}$

   (d) $[2x_1, 2x_2, \cdots, 2x_n]^T$

   (e) $[2x_1, 2x_2, \cdots, 2x_n]^T = 2[x_1, x_2, \cdots, x_n]^T = 2\mathbf{x}$

   (f) $[w_1, w_2, \cdots, w_n]^T$

   (g) $\mathbf{w}$

   The most important ones here are (d), (e), (f) and (g).

4. $\nabla z = [2x, -6y]^T$, so the direction to take is $-\nabla z|_{x=-1,y=0} = [2, 0]^T$.

part (a). Then

$$\begin{aligned}
\frac{d(\mathbf{x}^T\mathbf{A}\mathbf{x})}{d\mathbf{x}} &= \begin{bmatrix} \frac{d(A_{11}x_1^2+(A_{12}+A_{21})x_1x_2+A_{22}x_2^2)}{dx_1} \\ \frac{d(A_{11}x_1^2+(A_{12}+A_{21})x_1x_2+A_{22}x_2^2)}{dx_2} \end{bmatrix} \\
&= \begin{bmatrix} 2A_{11}x_1 + (A_{12}+A_{21})x_2 \\ (A_{12}+A_{21})x_1 + 2A_{22}x_2 \end{bmatrix} \\
&= \begin{bmatrix} (A_{11}+A_{11})x_1 + (A_{12}+A_{21})x_2 \\ (A_{21}+A_{12})x_1 + (A_{22}+A_{22})x_2 \end{bmatrix} \\
&= \begin{bmatrix} A_{11}x_1 + A_{12}x_2 \\ A_{21}x_1 + A_{22}x_2 \end{bmatrix} + \begin{bmatrix} A_{11}x_1 + A_{21}x_2 \\ A_{12}x_1 + A_{22}x_2 \end{bmatrix} \\
&= \begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} + \begin{bmatrix} A_{11} & A_{21} \\ A_{12} & A_{22} \end{bmatrix} \begin{bmatrix} x_1 \\ x_2 \end{bmatrix} \\
&= \mathbf{A}\mathbf{x} + \mathbf{A}^T\mathbf{x} \\
&= (\mathbf{A}+\mathbf{A}^T)\mathbf{x}.
\end{aligned}$$

(d)

$$\begin{aligned}
\mathbf{x}^T\mathbf{A}\mathbf{x} &= \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{n1} & \cdots & A_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \\
&= \begin{bmatrix} x_1 & \cdots & x_n \end{bmatrix} \begin{bmatrix} \sum_{j=1}^n A_{1j}x_j \\ \vdots \\ \sum_{j=1}^n A_{nj}x_j \end{bmatrix} \\
&= \sum_{i=1}^n x_i \left(\sum_{j=1}^n A_{ij}x_j \right).
\end{aligned}$$

Now we will determine which terms contain the factor x_k. This is to find $d(\mathbf{x}^T\mathbf{A}\mathbf{x})/dx_k$, which will be the k-th element of $d(\mathbf{x}^T\mathbf{A}\mathbf{x})/d\mathbf{x}$. x_k can appear when $i = k$ and/or $j = k$. When $i = k$, we have terms $\sum_{j=1}^n A_{kj}x_jx_k$, and when $j = k$, we have terms

$\sum_{i=1}^{n} A_{ik}x_i x_k$. Therefore

$$\begin{aligned}
\frac{d(\mathbf{x}^T\mathbf{A}\mathbf{x})}{dx_k} &= \frac{d}{dx_k}\left[\left(\sum_{i=1}^{n} A_{ik}x_i + \sum_{j=1}^{n} A_{kj}x_j\right)x_k - A_{kk}x_k^2\right] \quad (\because i=j=k \text{ counted twice}) \\
&= \frac{d}{dx_k}\left[\left(\sum_{i\neq k} A_{ik}x_i + \sum_{j\neq k} A_{kj}x_j\right)x_k + 2A_{kk}x_k^2 - A_{kk}x_k^2\right] \\
&= \left(\sum_{i\neq k} A_{ik}x_i + \sum_{j\neq k} A_{kj}x_j\right) + 2A_{kk}x_k \\
&= \sum_{i=1}^{n} A_{ik}x_i + \sum_{j=1}^{n} A_{kj}x_j \\
&= [\mathbf{A}^T\mathbf{x}]_k + [\mathbf{A}\mathbf{x}]_k \\
&= [(\mathbf{A}+\mathbf{A}^T)\mathbf{x}]_k.
\end{aligned}$$

As this is the k-th element of $d(\mathbf{x}^T\mathbf{A}\mathbf{x})/d\mathbf{x}$, it follows that

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

(e) $\mathbf{A}^T = \mathbf{A}$, so the result is simplified to $(\mathbf{A}+\mathbf{A}^T)\mathbf{x} = 2\mathbf{A}\mathbf{x}$.

4. Say $u(\mathbf{x}) = [u_1, \cdots, u_m]^T$ and $v(\mathbf{x}) = [v_1, \cdots, v_m]^T$. To show the two matrices are equivalent, it suffices to show that the (i,j)-th elements are the same for any i and j.

   (a) LHS: $u(\mathbf{x})+v(\mathbf{x}) = [u_1+v_1, \cdots, u_m+v_m]^T$, so the (i,j)-th element of $d(u+v)/d\mathbf{x}$ is $d((u+v)_j)/dx_i = d(u_j+v_j)/dx_i$. Because these are all scalars, using the scalar differentiation linearity, $d(u_j+v_j)/dx_i = du_j/dx_i + dv_j/dx_i$.
   RHS: The (i,j)-th element of $du/d\mathbf{x}$ is du_j/dx_i, and (i,j)-th element of $dv/d\mathbf{x}$ is dv_j/dx_i. Therefore, the (i,j)-th element of $du/d\mathbf{x} + dv/d\mathbf{x}$ is $du_j/dx_i + dv_j/dx_i$. The LHS and the RHS have the same (i,j)-th element.

   (b) $au(\mathbf{x}) = [au_1, \cdots, au_m]^T$. The (i,j)-th element of $d(au)/d\mathbf{x}$ is $d(au_j)/dx_i$, and again by the scalar differentiation linearity, $a \cdot du_j/dx_i$. The (i,j)-th element of $a \cdot du/d\mathbf{x}$ is $a \cdot du_j/dx_i$. The LHS and the RHS have the same (i,j)-th element.

5. Expanding everything (we have not derived the product rule, so we cannot take the derivative of this as is), we first have

$$\begin{aligned}
J(\boldsymbol{\theta}) &= \frac{1}{N}(\mathbf{X}\boldsymbol{\theta}-\mathbf{y})^T(\mathbf{X}\boldsymbol{\theta}-\mathbf{y}) \\
&= \frac{1}{N}(\boldsymbol{\theta}^T\mathbf{X}^T-\mathbf{y}^T)(\mathbf{X}\boldsymbol{\theta}-\mathbf{y}) \\
&= \frac{1}{N}(\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} - \boldsymbol{\theta}^T\mathbf{X}^T\mathbf{y} - \mathbf{y}^T\mathbf{X}\boldsymbol{\theta} + \mathbf{y}^T\mathbf{y}).
\end{aligned}$$

Observe that $d(\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta})/d\boldsymbol{\theta} = 2\mathbf{X}^T\mathbf{X}\boldsymbol{\theta}$, since $\mathbf{X}^T\mathbf{X}$ is a symmetric matrix and this follows directly from the previous problem. Also using the properties we have derived, $d(\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{y})/d\boldsymbol{\theta} = \mathbf{X}^T\mathbf{y}$ and $d(\mathbf{y}^T\mathbf{X}\boldsymbol{\theta})/d\boldsymbol{\theta} = \mathbf{X}^T\mathbf{y}$. $\mathbf{y}^T\mathbf{y}$ does not depend on $\boldsymbol{\theta}$, so $d(\mathbf{y}^T\mathbf{y})/d\boldsymbol{\theta} = \mathbf{0}$. Therefore, we have

$$\begin{aligned}\frac{dJ(\boldsymbol{\theta})}{d\boldsymbol{\theta}} &= \frac{1}{N}\left(\frac{d(\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{X}\boldsymbol{\theta})}{d\boldsymbol{\theta}} - \frac{d(\boldsymbol{\theta}^T\mathbf{X}^T\mathbf{y})}{d\boldsymbol{\theta}} - \frac{d(\mathbf{y}^T\mathbf{X}\boldsymbol{\theta})}{d\boldsymbol{\theta}} + \frac{d(\mathbf{y}^T\mathbf{y})}{d\boldsymbol{\theta}}\right) \\ &= \frac{1}{N}(2\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} - \mathbf{X}^T\mathbf{y} - \mathbf{X}^T\mathbf{y} + \mathbf{0}) \\ &= \frac{2}{N}(\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} - \mathbf{X}^T\mathbf{y}).\end{aligned}$$

Solving $dJ(\boldsymbol{\theta})/d\boldsymbol{\theta} = 0$ is therefore equivalent to solving $\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} - \mathbf{X}^T\mathbf{y} = 0$. Assuming that $\mathbf{X}^T\mathbf{X}$ is invertible, we have

$$\begin{aligned}&\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} - \mathbf{X}^T\mathbf{y} = 0 \\ \Rightarrow\ &\mathbf{X}^T\mathbf{X}\boldsymbol{\theta} = \mathbf{X}^T\mathbf{y} \\ \Rightarrow\ &\boldsymbol{\theta} = (\mathbf{X}^T\mathbf{X})^{-1}\mathbf{X}^T\mathbf{y}.\end{aligned}$$

6. df/dg and dg/dx are both scalars, and scalar multiplication is commutative. Therefore, we can safely swap the order of the multiplication and say $(df/dg)(dg/dx) = (dg/dx)(df/dx)$.

   On the other hand, $d\mathbf{z}/d\mathbf{y}$ and $d\mathbf{y}/d\mathbf{x}$ are both matrices, and matrix multiplication is *not* commutative. Therefore, we cannot swap the order of the multiplication. Specifically, $d\mathbf{z}/d\mathbf{x} \neq (d\mathbf{z}/d\mathbf{y})(d\mathbf{y}/d\mathbf{x})$; this multiplication is not even defined when the dimensions of $\mathbf{x}$ and $\mathbf{z}$ do not match.

7. (a)

$$\begin{aligned}\frac{\partial f}{\partial u} &= (u - v + 1)e^u, & \frac{\partial f}{\partial v} &= -e^u, \\ \frac{\partial u}{\partial x} &= y, & \frac{\partial u}{\partial y} &= x, \\ \frac{\partial v}{\partial x} &= 2x, & \frac{\partial v}{\partial y} &= -2y.\end{aligned}$$

   By the chain rule,

$$\begin{aligned}\frac{\partial f}{\partial x} = \frac{\partial f}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial x} &= (u - v + 1)e^u y - 2xe^u \\ &= (xy - x^2 + y^2 + 1)ye^{xy} - 2xe^{xy}, \\ \frac{\partial f}{\partial y} = \frac{\partial f}{\partial u}\frac{\partial u}{\partial y} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial y} &= (u - v + 1)e^u x + 2ye^u \\ &= (xy - x^2 + y^2 + 1)xe^{xy} + 2ye^{xy}.\end{aligned}$$

(b)

$$\begin{aligned}
&\frac{\partial f}{\partial u} = \log v + \frac{v}{u}, && \frac{\partial f}{\partial v} = \log u + \frac{u}{v},\\
&\frac{\partial u}{\partial x} = \frac{1}{2}, && \frac{\partial u}{\partial y} = -\frac{2}{y^2},\\
&\frac{\partial v}{\partial x} = e^y, && \frac{\partial v}{\partial y} = xe^y.
\end{aligned}$$

By the chain rule,

$$\begin{aligned}
\frac{\partial f}{\partial x} = \frac{\partial f}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial x} &= \frac{1}{2}\left(\log v + \frac{v}{u}\right) + \left(\log u + \frac{u}{v}\right)e^y\\
&= \frac{1}{2}\left(\log x + y + \frac{xe^y}{\frac{x}{2}+\frac{2}{y}}\right) + \left(\log\left(\frac{x}{2}+\frac{2}{y}\right) + \frac{\frac{x}{2}+\frac{2}{y}}{xe^y}\right)e^y,\\
\frac{\partial f}{\partial y} = \frac{\partial f}{\partial u}\frac{\partial u}{\partial y} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial y} &= -\frac{2}{y^2}\left(\log v + \frac{v}{u}\right) + \left(\log u + \frac{u}{v}\right)xe^y\\
&= -\frac{2}{y^2}\left(\log x + y + \frac{xe^y}{\frac{x}{2}+\frac{2}{y}}\right) + \left(\log\left(\frac{x}{2}+\frac{2}{y}\right) + \frac{\frac{x}{2}+\frac{2}{y}}{xe^y}\right)xe^y.
\end{aligned}$$

(c)

$$\begin{aligned}
&\frac{\partial f}{\partial u} = \log v, && \frac{\partial f}{\partial v} = \frac{u}{v},\\
&\frac{\partial u}{\partial x} = \sin y + y\cos x, && \frac{\partial u}{\partial y} = x\cos y + \sin x,\\
&\frac{\partial v}{\partial x} = \cos y - y\sin x, && \frac{\partial v}{\partial y} = -x\sin y + \cos x.
\end{aligned}$$

By the chain rule,

$$\begin{aligned}
\frac{\partial f}{\partial x} &= \frac{\partial f}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial x}\\
&= (\sin y + y\cos x)\log v + (\cos y - y\sin x)\frac{u}{v}\\
&= (\sin y + y\cos x)\log(x\cos y + y\cos x) + (\cos y - y\sin x)\frac{x\sin y + y\sin x}{x\cos y + y\cos x},\\
\frac{\partial f}{\partial y} &= \frac{\partial f}{\partial u}\frac{\partial u}{\partial y} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial y}\\
&= (x\cos y + \sin x)\log v + (\cos x - x\sin y)\frac{u}{v}\\
&= (x\cos y + \sin x)\log(x\cos y + y\cos x) + (\cos x - x\sin y)\frac{x\sin y + y\sin x}{x\cos y + y\cos x}.
\end{aligned}$$

(d)

$$\frac{\partial f}{\partial u} = \frac{v^2+1}{(1-uv)^2}, \qquad \frac{\partial f}{\partial v} = \frac{u^2+1}{(1-uv)^2},$$
$$\frac{\partial u}{\partial x} = \frac{1}{2}\sec^2\frac{x+y}{2}, \qquad \frac{\partial u}{\partial y} = \frac{1}{2}\sec^2\frac{x+y}{2},$$
$$\frac{\partial v}{\partial x} = \frac{1}{2}\sec^2\frac{x-y}{2}, \qquad \frac{\partial v}{\partial y} = -\frac{1}{2}\sec^2\frac{x-y}{2},$$

where $\sec x = 1/\cos x$. By the chain rule and using that $1+\tan^2 x = \sec^2 x$,

$$\begin{aligned}
\frac{\partial f}{\partial x} &= \frac{\partial f}{\partial u}\frac{\partial u}{\partial x} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial x} \\
&= \frac{v^2+1}{2(1-uv)^2}\sec^2\frac{x+y}{2} + \frac{u^2+1}{2(1-uv)^2}\sec^2\frac{x-y}{2} \\
&= \frac{1+\tan^2\frac{x-y}{2}}{2\left(1-\tan\frac{x+y}{2}\tan\frac{x-y}{2}\right)^2}\sec^2\frac{x+y}{2} + \frac{1+\tan^2\frac{x+y}{2}}{2\left(1-\tan\frac{x+y}{2}\tan\frac{x-y}{2}\right)^2}\sec^2\frac{x-y}{2} \\
&= \frac{\sec^2\frac{x-y}{2}\sec^2\frac{x+y}{2}}{2\left(1-\tan\frac{x+y}{2}\tan\frac{x-y}{2}\right)^2} + \frac{\sec^2\frac{x+y}{2}\sec^2\frac{x-y}{2}}{2\left(1-\tan\frac{x+y}{2}\tan\frac{x-y}{2}\right)^2} \\
&= \frac{\sec^2\frac{x-y}{2}\sec^2\frac{x+y}{2}}{\left(1-\tan\frac{x+y}{2}\tan\frac{x-y}{2}\right)^2} \\
&= \frac{1}{\left(\cos\frac{x+y}{2}\cos\frac{x-y}{2} - \sin\frac{x+y}{2}\sin\frac{x-y}{2}\right)^2} \\
&= \frac{1}{\cos^2\left(\frac{x+y}{2}+\frac{x-y}{2}\right)} \\
&= \frac{1}{\cos^2 x} = \sec^2 x, \\
\frac{\partial f}{\partial y} &= \frac{\partial f}{\partial u}\frac{\partial u}{\partial y} + \frac{\partial f}{\partial v}\frac{\partial v}{\partial y} \\
&= \frac{v^2+1}{2(1-uv)^2}\sec^2\frac{x+y}{2} - \frac{u^2+1}{2(1-uv)^2}\sec^2\frac{x-y}{2} \\
&= \frac{1+\tan^2\frac{x-y}{2}}{2\left(1-\tan\frac{x+y}{2}\tan\frac{x-y}{2}\right)^2}\sec^2\frac{x+y}{2} - \frac{1+\tan^2\frac{x+y}{2}}{2\left(1-\tan\frac{x+y}{2}\tan\frac{x-y}{2}\right)^2}\sec^2\frac{x-y}{2} \\
&= \frac{\sec^2\frac{x-y}{2}\sec^2\frac{x+y}{2}}{2\left(1-\tan\frac{x+y}{2}\tan\frac{x-y}{2}\right)^2} - \frac{\sec^2\frac{x+y}{2}\sec^2\frac{x-y}{2}}{2\left(1-\tan\frac{x+y}{2}\tan\frac{x-y}{2}\right)^2} \\
&= 0.
\end{aligned}$$

Did you notice that $(u+v)/(1-uv)$ with $u = \tan\frac{x+y}{2}$ and $v = \tan\frac{x-y}{2}$ is simply

$$\frac{\tan\frac{x+y}{2}+\tan\frac{x-y}{2}}{1-\tan\frac{x+y}{2}\tan\frac{x-y}{2}} = \tan\left(\frac{x+y}{2}+\frac{x-y}{2}\right) = \tan x,$$

so $\partial f/\partial x = \sec^2 x$ and $\partial f/\partial y = 0$?

8. (a) By the chain rule,

$$\frac{dl}{db_1} = \frac{d\hat{y}}{db_1}\frac{dl}{d\hat{y}}.$$

You may be more familiar with the form that has the multiplication in the reversed order: $dl/db_1 = (dl/d\hat{y})(d\hat{y}/db_1)$. Switching the order does not matter as these are scalars, but this order is more consistent with the vector version.

Now $\hat{y} = h(\mathbf{b}) = h(b_1, b_2)$. Therefore, $d\hat{y}/db_1$ is simply $\partial h/\partial b_1$, and the answer is

$$\frac{dl}{db_1} = \frac{dh}{db_1}\frac{dl}{d\hat{y}}.$$

Here we are writing that $d\hat{y}/db_1 = dh(\mathbf{b})/db_1 = dh/db_1$. We are also slightly abusing the notation here (as we always have) and writing ∂ as d.

(b) Similarly, $dl/db_2 = (dh/db_2)(dl/d\hat{y})$.

$$\begin{aligned}
\frac{dl}{d\mathbf{b}} &= \left[\frac{dl}{db_1}, \frac{dl}{db_2}\right]^T \\
&= \left[\frac{dh}{db_1}\frac{dl}{d\hat{y}}, \frac{dh}{db_2}\frac{dl}{d\hat{y}}\right]^T \\
&= \left[\frac{dh}{db_1}, \frac{dh}{db_2}\right]^T \frac{dl}{d\hat{y}} \\
&= \frac{dh}{d\mathbf{b}}\frac{dl}{d\hat{y}}.
\end{aligned}$$

(c) Isolating only a_2 from the diagram, we know that the chain rule should be

$$\frac{dl}{da_2} = \frac{db_1}{da_2}\frac{dl}{db_1} + \frac{db_2}{da_2}\frac{dl}{db_2}.$$

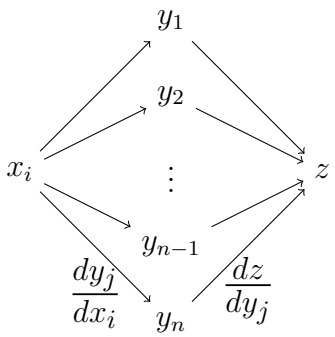
(d) By the chain rule, we have

$$\frac{dl}{d\mathbf{a}} = \frac{d\mathbf{b}}{d\mathbf{a}}\frac{dl}{d\mathbf{b}}.$$

Now notice that $\mathbf{b} = g(\mathbf{a})$, so $d\mathbf{b}/d\mathbf{a}$ is simply $dg/d\mathbf{a}$. Therefore, the answer is

$$\frac{dl}{d\mathbf{a}} = \frac{dg}{d\mathbf{a}}\frac{dl}{d\mathbf{b}}.$$

9. (a) Again, we draw the diagram:



The result directly follows from the diagram. Each $x_i \to y_j \to z$ path gives $(dy_j/dx_i)(dz/dy_j)$, and we add it for all possible values of j.

(b) This is defined in the numerator layout. $\nabla_{\mathbf{x}} z \in \mathbb{R}^m$ and $\nabla_{\mathbf{x}} y \in \mathbb{R}^n$, so $(\partial \mathbf{y}/\partial \mathbf{x})^T$ has to be $\mathbb{R}^{m \times n}$, which means $\partial \mathbf{y}/\partial \mathbf{x}$ itself is $\mathbb{R}^{n \times m}$. This is the shape when we use the numerator-layout notation.

(c) We have already derived the version using the denominator-layout notation, which was $\nabla_{\mathbf{x}} z = (\partial \mathbf{y}/\partial \mathbf{x}) \nabla_{\mathbf{y}} z$. Note that with our definition, $\nabla_{\mathbf{x}} z = \partial z/\partial \mathbf{x}$ and $\nabla_{\mathbf{y}} z = \partial z/\partial \mathbf{y}$. Therefore, this is exactly the same as the chain rule we have derived: $\partial z/\partial \mathbf{x} = (\partial \mathbf{y}/\partial \mathbf{x})(\partial z/\partial \mathbf{y})$.

Now we consider the numerator-layout notation. In this convention, all definitions are transposed from the ones we used so far, so the derivation for the general case ($\mathbf{x} \in \mathbb{R}^p$, $\mathbf{y} \in \mathbb{R}^r$, $\mathbf{z} \in \mathbb{R}^n$ where $\mathbf{z}$ is a function of $\mathbf{y}$, and $\mathbf{y}$ is a function of $\mathbf{x}$) becomes:

$$
\begin{aligned}
\frac{d\mathbf{z}}{d\mathbf{x}} &= \begin{bmatrix} dz_1/dx_1 & dz_1/dx_2 & \cdots & dz_1/dx_p \\ dz_2/dx_1 & dz_2/dx_2 & \cdots & dz_2/dx_p \\ \vdots & & \ddots & \vdots \\ dz_n/dx_1 & dz_n/dx_2 & \cdots & dz_n/dx_p \end{bmatrix} \in \mathbb{R}^{n \times p} \\
&= \begin{bmatrix} \sum_{k=1}^r \frac{dz_1}{dy_k}\frac{dy_k}{dx_1} & \sum_{k=1}^r \frac{dz_1}{dy_k}\frac{dy_k}{dx_2} & \cdots & \sum_{k=1}^r \frac{dz_1}{dy_k}\frac{dy_k}{dx_n} \\ \sum_{k=1}^r \frac{dz_2}{dy_k}\frac{dy_k}{dx_1} & \sum_{k=1}^r \frac{dz_2}{dy_k}\frac{dy_k}{dx_2} & \cdots & \sum_{k=1}^r \frac{dz_2}{dy_k}\frac{dy_k}{dx_n} \\ \vdots & & \ddots & \vdots \\ \sum_{k=1}^r \frac{dz_p}{dy_k}\frac{dy_k}{dx_1} & \sum_{k=1}^r \frac{dz_p}{dy_k}\frac{dy_k}{dx_2} & \cdots & \sum_{k=1}^r \frac{dz_p}{dy_k}\frac{dy_k}{dx_n} \end{bmatrix} \\
&= \begin{bmatrix} dz_1/dy_1 & dz_1/dy_2 & \cdots & dz_1/dy_r \\ dz_2/dy_1 & dz_2/dy_2 & \cdots & dz_2/dy_r \\ \vdots & & \ddots & \vdots \\ dz_n/dy_1 & dz_n/dy_2 & \cdots & dz_n/dy_r \end{bmatrix} \begin{bmatrix} dy_1/dx_1 & dy_1/dx_2 & \cdots & dy_1/dx_p \\ dy_2/dx_1 & dy_2/dx_2 & \cdots & dy_2/dx_p \\ \vdots & & \ddots & \vdots \\ dy_r/dx_1 & dy_r/dx_2 & \cdots & dy_r/dx_p \end{bmatrix} \\
&= \frac{d\mathbf{z}}{d\mathbf{y}} \frac{d\mathbf{y}}{d\mathbf{x}}.
\end{aligned}
$$

For this problem, z is a scalar, which can be considered as a $\mathbf{z} \in \mathbb{R}^1$ vector. Therefore, we have $dz/d\mathbf{x} = (dz/d\mathbf{y})(d\mathbf{y}/d\mathbf{x})$. Compare this to Eq. (G6.46). Notice that $\nabla_{\mathbf{x}} z = (dz/d\mathbf{x})^T$ and $\nabla_{\mathbf{y}} z = (dz/d\mathbf{y})^T$ with the transposed definitions. Taking the transpose of both sides of $dz/d\mathbf{x} = (dz/d\mathbf{y})(d\mathbf{y}/d\mathbf{x})$, we obtain $(dz/d\mathbf{x})^T = (d\mathbf{y}/d\mathbf{x})^T (dz/d\mathbf{y})^T$, or equivalently $\nabla_{\mathbf{x}} z = (d\mathbf{y}/d\mathbf{x})^T \nabla_{\mathbf{y}} z$, which is identical to Eq. (G6.46).

10. (a) By the exactly same derivation from the previous problem,

$$\frac{d\mathbf{z}}{d\mathbf{x}} = \frac{d\mathbf{z}}{d\mathbf{y}} \frac{d\mathbf{y}}{d\mathbf{x}}.$$

(b) Keeping in mind that $dz/d\mathbf{x}$ is equivalent to the gradient (column vector), we

have for $\mathbf{x} \in \mathbb{R}^p$ and $\mathbf{y} \in \mathbb{R}^r$,

$$\begin{aligned}
\left(\frac{dz}{d\mathbf{x}}\right)^T &= \begin{bmatrix} \dfrac{dz}{dx_1} & \cdots & \dfrac{dz}{dx_p} \end{bmatrix} \\
&= \begin{bmatrix} \displaystyle\sum_{k=1}^{r} \frac{dz}{dy_k}\frac{dy_k}{dx_1} & \cdots & \displaystyle\sum_{k=1}^{r} \frac{dz}{dy_k}\frac{dy_k}{dx_p} \end{bmatrix} \\
&= \begin{bmatrix} \dfrac{dz}{dy_1} & \cdots & \dfrac{dz}{dy_r} \end{bmatrix} \begin{bmatrix} \dfrac{dy_1}{dx_1} & \dfrac{dy_1}{dx_2} & \cdots & \dfrac{dy_1}{dx_p} \\ \dfrac{dy_2}{dx_1} & \dfrac{dy_2}{dx_2} & \cdots & \dfrac{dy_2}{dx_p} \\ \vdots & \ddots & & \vdots \\ \dfrac{dy_r}{dx_1} & \dfrac{dy_r}{dx_2} & \cdots & \dfrac{dy_r}{dx_p} \end{bmatrix} \\
&= \left(\frac{dz}{d\mathbf{y}}\right)^T \frac{d\mathbf{y}}{d\mathbf{x}}.
\end{aligned}$$

Another way of thinking about this is that if $dz/d\mathbf{x}$ were transposed as well, we can use the chain rule from the numerator layout. Therefore, we simply transpose $dz/d\mathbf{x}$ and $dz/d\mathbf{y}$, then apply the same rule.

(c) Taking the transpose of the result from part (b), we have

$$\frac{dz}{d\mathbf{x}} = \left(\frac{d\mathbf{y}}{d\mathbf{x}}\right)^T \frac{dz}{d\mathbf{y}},$$

which is equivalent to Eq. (G6.46) (in this definition, $\nabla_{\mathbf{v}} z = dz/d\mathbf{v}$). Although the derived chain rule is the same, the reasoning is different. For Question 9, we can simply derive the general chain rule for $d\mathbf{z}/d\mathbf{x}$, then argue that a scalar is a one-dimensional vector and therefore is just a special case of the general case. This works because $dy/d\mathbf{x} = d\mathbf{y}/d\mathbf{x}$ when we interpret $y \in \mathbb{R}$ as $\mathbf{y} = [y]^T \in \mathbb{R}^1$. However, this does not hold under the new definition, which means we have to derive the chain rule specifically for $dz/d\mathbf{x}$ (or appropriately transpose the result from a different set of definitions). The caveat is that we have to be careful when scalar derivatives and vector derivatives cross paths.

4.3 Section 3

1. (a) $\mathbf{u}^T\mathbf{v}$ is a scalar, so the expected shape is the same as that of $\mathbf{x}$. $d(\mathbf{u}^T\mathbf{v})/d\mathbf{x} \in \mathbb{R}^m$.

   (b) $d\mathbf{u}/d\mathbf{x}, d\mathbf{v}/d\mathbf{x} \in \mathbb{R}^{m\times n}$ and $\mathbf{u}, \mathbf{v} \in \mathbb{R}^m$. Options traversed from left to right, then from up to down.

   - $(d\mathbf{u}/d\mathbf{x})(d\mathbf{v}/d\mathbf{x})$ is not defined.
   - $(d\mathbf{v}/d\mathbf{x})(d\mathbf{u}/d\mathbf{x})$ is not defined.
   - $\mathbf{u}(d\mathbf{u}/d\mathbf{x})$ is not defined.
   - $\mathbf{v}(d\mathbf{u}/d\mathbf{x})$ is not defined.
   - $\mathbf{u}(d\mathbf{v}/d\mathbf{x})$ is not defined (faintly surprising that this is not the answer?).
   - $\mathbf{u}(d\mathbf{v}/d\mathbf{x})$ is not defined.
   - $(d\mathbf{u}/d\mathbf{x})\mathbf{v}$ is defined and in $\mathbb{R}^m$. $(d\mathbf{v}/d\mathbf{x})\mathbf{u}$ is also defined and in $\mathbb{R}^m$. Adding the two gives $\mathbb{R}^m$, which is the expected shape. This can potentially be the answer.
   - $(d\mathbf{u}/d\mathbf{x})\mathbf{u}$ is defined and in $\mathbb{R}^m$. $(d\mathbf{v}/d\mathbf{x})\mathbf{v}$ is also defined and in $\mathbb{R}^m$. Adding the two gives $\mathbb{R}^m$, which is the expected shape. This can potentially be the answer.

   Therefore, the last two options can potentially be the answer. This shows the limitation of shape matching; sometimes you are left with multiple possible options.

   (c) The i-th element of $d(\mathbf{u}^T\mathbf{v})/d\mathbf{x}$ is $d(\mathbf{u}^T\mathbf{v})/dx_i$. $\mathbf{u}^T\mathbf{v} = u_1v_1 + \cdots + u_nv_n$. Applying the scalar chain rule,

$$\begin{aligned}
\frac{d(\mathbf{u}^T\mathbf{v})}{dx_i} &= \frac{d}{dx_i}(u_1v_1 + \cdots + u_nv_n) \\
&= \frac{d}{dx_i}u_1v_1 + \cdots + \frac{d}{dx_i}u_nv_n \\
&= \left(\frac{du_1}{dx_i}v_1 + u_1\frac{dv_1}{dx_i}\right) + \cdots + \left(\frac{du_n}{dx_i}v_n + u_n\frac{dv_n}{dx_i}\right) \\
&= \left(\frac{du_1}{dx_i}v_1 + \cdots + \frac{du_n}{dx_i}v_n\right) + \left(u_1\frac{dv_1}{dx_i} + \cdots + u_n\frac{dv_n}{dx_i}\right).
\end{aligned}$$

   (d) $d(\mathbf{u}^T\mathbf{v})/dx_i$ is the sum of two vector inner products. Observe that

$$\begin{aligned}
\frac{du_1}{dx_i}v_1 + \cdots + \frac{du_n}{dx_i}v_n &= \frac{d\mathbf{u}}{dx_i}\mathbf{v}, \\
u_1\frac{dv_1}{dx_i} + \cdots + u_n\frac{dv_n}{dx_i} &= \frac{d\mathbf{v}}{dx_i}\mathbf{u}.
\end{aligned}$$

   Recall that $d\mathbf{u}/dx_i$ and $d\mathbf{v}/dx_i$ are row vectors, so the multiplication results are scalars. Generalizing this to the entire vector, we have

$$\frac{d(\mathbf{u}^T\mathbf{v})}{d\mathbf{x}} = \frac{d\mathbf{u}}{d\mathbf{x}}\mathbf{v} + \frac{d\mathbf{v}}{d\mathbf{x}}\mathbf{u}.$$

2. (a) $d\mathbf{y}/d\mathbf{x}$ is the derivative of a vector with respect to a vector. Therefore, the derivative is a matrix. The (i,j)-th element of $d\mathbf{y}/d\mathbf{x}$ is dy_j/dx_i. Notice that this is nonzero only where $i=j$; $dy_i/dx_i = df(x_i)/dx_i$, and zero everywhere else. Therefore,

$$\begin{aligned}\frac{d\mathbf{y}}{d\mathbf{x}} &= \begin{bmatrix} f'(x_1) & & & \\ & f'(x_2) & \mathbf{0} & \\ & & \ddots & \\ & \mathbf{0} & & f'(x_{n-1}) \\ & & & & f'(x_n) \end{bmatrix} \\ &= \operatorname{diag}(f'(x_1), \cdots, f'(x_n)) \\ &= \operatorname{diag}(f'(\mathbf{x})).\end{aligned}$$

where f' is the derivative of f and is applied element-wise to $\mathbf{x}$.

diag is a function you will see often, which creates a matrix with the argument as the main diagonal.

(b) Again, the derivative is a matrix. The (i,j)-th element of $d\mathbf{z}/d\mathbf{x}$ is dz_j/dx_i. This can be nonzero only where $i=j$; $dz_i/dx_i = y_i$, and zero everywhere else. Therefore,

$$\frac{d\mathbf{z}}{d\mathbf{x}} = \operatorname{diag}(\mathbf{y}).$$

$\odot$, the element-wise product, is also called the Hadamard product.

(c) This generalizes the previous part of this question, and now $dz_j/dx_i = df(x_j, y_j)/dx_i$. Again, this is only nonzero along the main diagonal of the matrix. Therefore,

$$\frac{d\mathbf{z}}{d\mathbf{x}} = \operatorname{diag}\left(\frac{\partial f}{\partial x_1}, \cdots, \frac{\partial f}{\partial x_n}\right).$$

3. (a) $y_i = \sum_{p=1}^{n} X_{ip} v_p$. This means that y_i depends only on $X_{i1}, \cdots, X_{in}$. Therefore, dy_i/dX_{jk} is necessarily zero where $i \neq j$.

(b) We consider dy_i/dX_{jk} only when $i=j$. Then $dy_i/dX_{jk} = dy_i/dX_{ik} = v_k$.

(c) We are essentially computing only the values of dy_i/dX_{ik}. There are only two free variables here, namely i and k. This means the information can be represented as a 2D matrix; say $\mathbf{M} \in \mathbb{R}^{m \times n}$ where $M_{ij} = dy_i/dX_{ij} = v_j$.

(d) The matrix constructed in (c) is

$$\begin{bmatrix} v_1 & \cdots & v_n \\ \vdots & \ddots & \vdots \\ v_1 & \cdots & v_n \end{bmatrix}.$$

All rows of the matrix are the same, so they are redundant. No matter how you construct the matrix in (c), M_{ij} having two free variables and v_i having only one means that one of the free variables of M_{ij} is unnecessary.

(e) We delete all rows except one, and say the answer is $[v_1, \cdots, v_n]^T = \mathbf{v}^T$.

(f) Say $\mathbf{X} = \mathbf{x}^T = [x_1, \cdots, x_n]^T$. This is now the derivative of a scalar with respect to a (row) vector, and by the definition we have, this is $\mathbf{v}^T = [v_1, \cdots, v_n]^T$. Defining the answer to be $\mathbf{v}^T$ is a smooth generalization.

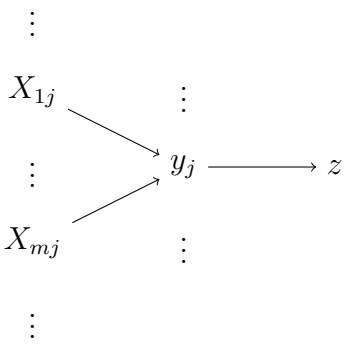
4. (a) $d\mathbf{y}/d\mathbf{X} = d(\mathbf{v}^T\mathbf{X})/d\mathbf{X} = \mathbf{v} \in \mathbb{R}^{m\times 1}$, $dz/d\mathbf{X} \in \mathbb{R}^{m\times n}$, and $dz/d\mathbf{y} \in \mathbb{R}^{1\times n}$. We expect the answer to be

$$\frac{dz}{d\mathbf{X}} = \frac{d\mathbf{y}}{d\mathbf{X}}\frac{dz}{d\mathbf{y}}.$$

(b) We start by considering the (i, j)-th element of $dz/d\mathbf{X}$, dz/dX_{ij}. Recall that $\mathbf{y}$ is a row vector; we denote its k-th element as y_k. Now we apply the scalar chain rule. Drawing the usual diagram for the chain rule, we first have:

$$\mathbf{X} \longrightarrow \mathbf{y} \longrightarrow z$$

As $\mathbf{y} = \mathbf{v}^T\mathbf{X}$, $y_j = \sum_{k=1}^m v_k X_{kj}$. This means that y_j only depends on $X_{1j}, \cdots, X_{mj}$, so the diagram can be specified as:



Finally, considering only one element of $\mathbf{X}$, namely X_{ij}, the diagram is simplified as:

$$X_{ij} \longrightarrow y_j \longrightarrow z$$

Applying the scalar chain rule here, we have

$$\frac{dz}{dX_{ij}} = \frac{dz}{dy_j}\frac{dy_j}{dX_{ij}}.$$

Since we already know that $y_j = \sum_{k=1}^m v_k X_{kj}$, we have that $dy_j/dX_{ij} = v_i$, and therefore

$$\frac{dz}{dX_{ij}} = \frac{dz}{dy_j}v_i = v_i\frac{dz}{dy_j}.$$

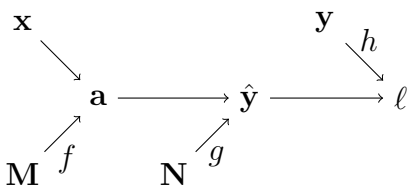
Recall that for vectors $\mathbf{a}$ and $\mathbf{b}$, $\mathbf{D} = \mathbf{a}\mathbf{b}^T$ is defined as $D_{ij} = a_i b_j$. Here with $\mathbf{a} = \mathbf{v}$ and $\mathbf{b}^T = dz/d\mathbf{y}$, we can conclude that

$$\frac{dz}{d\mathbf{X}} = \mathbf{v}\frac{dz}{d\mathbf{y}}.$$

Note that $\mathbf{b}^T$ (not just $\mathbf{b}$) should be matched with $dz/d\mathbf{y}$, because $dz/d\mathbf{y}$ is a row vector.

(c) Comparing the answers of (a) and (b), $d\mathbf{y}/d\mathbf{X} = \mathbf{v}$ as we have defined. It is valid to justify the definition this way as we have never relied on $d\mathbf{y}/d\mathbf{X} = \mathbf{v}$ to derive (b). However, note that this does not constitute as a formal proof of $d\mathbf{y}/d\mathbf{X} = \mathbf{v}$.

5. (a) This is a typical neural network update situation.



By the chain rule, we have

$$\frac{d\ell}{d\mathbf{N}} = \frac{d\ell}{d\hat{\mathbf{y}}}\frac{d\hat{\mathbf{y}}}{d\mathbf{N}}.$$

Note that this looks slightly different from what we had so far. This can be "derived" using shape matching, and we can also prove this formally. First observe that

$$\frac{d\ell}{dN_{ij}} = \frac{d\hat{y}_i}{dN_{ij}}\frac{d\ell}{d\hat{y}_i}.$$

An easy thought process to obtain this is to first consider N_{ij}. We know $\mathbf{N}$ will affect $\hat{\mathbf{y}}$, and we see that $\hat{\mathbf{y}} = g(\mathbf{Na})$. The k-th element of the vector $\hat{\mathbf{y}}$ is defined as $g(\sum_p N_{kp}a_p)$, and here we know that N_{ij} contributes to $\hat{y}_i$. Then $\hat{y}_i$ affects ℓ. Now we convert it back to vectors/matrices. Again, this matches $\mathbf{D} = \mathbf{vw}^T$ where $v_i = d\ell/d\hat{y}_i$, $w_j = d\hat{y}_i/dN_{ij}$, and therefore $\mathbf{v} = d\ell/d\hat{\mathbf{y}}$ and $\mathbf{w}^T = d\hat{\mathbf{y}}/d\mathbf{N}$ with $\mathbf{D} = d\ell/d\mathbf{N}$.

(b) From the scalar chain rule,

$$\begin{aligned}\frac{d\ell}{dN_{ij}} &= \frac{d\hat{y}_i}{dN_{ij}}\frac{d\ell}{d\hat{y}_i} \\ &= \frac{dg(\sum_{k=1}^p N_{ik}a_k)}{dN_{ij}}\frac{d\ell}{d\hat{y}_i} \\ &= a_j g'(\hat{y}_i)\frac{d\ell}{d\hat{y}_i}.\end{aligned}$$

We match this again with $\mathbf{D} = \mathbf{vw}^T$. $D_{ij} = v_i w_j$, so $v_i = g'(\hat{y}_i)(d\ell/d\hat{y}_i)$ and $w_j = a_j$ naturally works. Therefore, $\mathbf{w} = \mathbf{a}$. $\mathbf{v}$ is a little trickier; $v_i = g'(\hat{y}_i)(d\ell/d\hat{y}_i)$ can be interpreted as an element-wise product, which gives $\mathbf{v} = g'(\hat{\mathbf{y}}) \odot (d\ell/d\hat{\mathbf{y}})$ where g' is the derivative of g and is applied element-wise. Hence,

$$\frac{d\ell}{d\mathbf{N}} = \left(g'(\hat{\mathbf{y}}) \odot \frac{d\ell}{d\hat{\mathbf{y}}}\right)\mathbf{a}^T.$$

(c) Observe the diagram we obtain when we isolate only $\mathbf{M}$, $\mathbf{a}$ and ℓ, and only $\mathbf{N}$, $\hat{\mathbf{y}}$ and ℓ.

The diagrams and the definitions of $\mathbf{a}$ and $\hat{\mathbf{y}}$ imply that we can directly reuse the result from (a). Therefore,

$$\frac{d\ell}{d\mathbf{M}} = \frac{d\ell}{d\mathbf{a}}\frac{d\mathbf{a}}{d\mathbf{M}}.$$

Also isolating only $\mathbf{a}$, $\hat{\mathbf{y}}$, and ℓ, we also have

$$\mathbf{a} \longrightarrow \hat{\mathbf{y}} \longrightarrow \ell$$

which is the case we can apply the vector chain rule we have derived. This gives

$$\frac{d\ell}{d\mathbf{a}} = \frac{d\hat{\mathbf{y}}}{d\mathbf{a}}\frac{d\ell}{d\hat{\mathbf{y}}}.$$

We can thus conclude that

$$\frac{d\ell}{d\mathbf{M}} = \frac{d\hat{\mathbf{y}}}{d\mathbf{a}}\frac{d\ell}{d\hat{\mathbf{y}}}\frac{d\mathbf{a}}{d\mathbf{M}}.$$

Alternatively, we can derive this starting from one element again. Consider $d\ell/dM_{ij}$:

$$\begin{aligned}\frac{d\ell}{dM_{ij}} &= \frac{da_i}{dM_{ij}}\sum_{k=1}^{p}\frac{d\hat{y}_k}{da_i}\frac{d\ell}{d\hat{y}_k} \\ &= \left(\sum_{k=1}^{p}\frac{d\hat{y}_k}{da_i}\frac{d\ell}{d\hat{y}_k}\right)\frac{da_i}{dM_{ij}}.\end{aligned}$$

Inside the parentheses is the i-th element of $(d\mathbf{y}/d\mathbf{a})(d\ell/d\hat{\mathbf{y}})$, and da_i/dM_{ij} is the j-th element of $d\mathbf{a}/d\mathbf{M}$. Therefore, we have the same chain rule.

(d) We also reuse the result from (b) to acquire

$$\frac{d\ell}{d\mathbf{M}} = \left(f'(\mathbf{a}) \odot \frac{d\ell}{d\mathbf{a}}\right)\mathbf{x}^T.$$

Applying the vector chain rule to $d\ell/d\mathbf{a}$ again yields

$$\frac{d\ell}{d\mathbf{M}} = \left(f'(\mathbf{a}) \odot \frac{d\hat{\mathbf{y}}}{d\mathbf{a}}\frac{d\ell}{d\hat{\mathbf{y}}}\right)\mathbf{x}^T.$$

$d\hat{\mathbf{y}}/d\mathbf{a}$ can be directly evaluated, whose (i,j)-th element is $d\hat{y}_j/da_i = dg(\sum_k N_{jk}a_k)/da_i = g'(\hat{y}_j)N_{ji}$. Therefore, $d\hat{\mathbf{y}}/d\mathbf{a} = g'(\hat{\mathbf{y}}^T) \odot \mathbf{N}^T$. The element-wise multiplication here should be broadcasted (the j-th column is $g'(\hat{y}_j)\mathbf{N}^T_{:,j}$). With this, we finally have

$$\frac{d\ell}{d\mathbf{M}} = \left[f'(\mathbf{a}) \odot \left(g'(\hat{\mathbf{y}}^T) \odot \mathbf{N}^T\right)\frac{d\ell}{d\hat{\mathbf{y}}}\right]\mathbf{x}^T.$$

(e) We revise the definitions as follows for $i = 1, \cdots, B$:

$$\begin{aligned}\mathbf{a}_i &= f(\mathbf{M}\mathbf{x}_i),\\ \hat{\mathbf{y}}_i &= g(\mathbf{N}\mathbf{a}_i),\\ \ell_i &= h(\hat{\mathbf{y}}_i, \mathbf{y}_i).\end{aligned}$$

As stated in the question, the new loss value ℓ is $\ell = \frac{1}{B}\sum_{i=1}^{B} \ell_i$. We can evaluate $d\ell_i/d\mathbf{M}$ and $d\ell_i/d\mathbf{N}$ exactly the same way as how we did for the previous part of this question. The new gradients are simply

$$\begin{aligned}\frac{d\ell}{d\mathbf{M}} &= \frac{1}{B}\sum_{i=1}^{B} \frac{d\ell_i}{d\mathbf{M}},\\ \frac{d\ell}{d\mathbf{N}} &= \frac{1}{B}\sum_{i=1}^{B} \frac{d\ell_i}{d\mathbf{N}}\end{aligned}$$

by linearity.

6. $\alpha_t(j) = A_{jx_t}\sum_{k=1}^{J} \alpha_{t-1}(k)B_{kj}$ is one element of a vector (the j-th element of $\boldsymbol{\alpha}_t$). Visualizing $\sum_{k=1}^{J} \alpha_{t-1}(k)B_{kj}$, we have

$$\begin{bmatrix}\alpha_{t-1}(1) & \cdots & \alpha_{t-1}(J)\end{bmatrix}\begin{bmatrix} & B_{1j} & \\ \cdots & \vdots & \cdots \\ & B_{Jj} & \end{bmatrix}$$

which can be expressed as $\boldsymbol{\alpha}_{t-1}^T\mathbf{B}_{:,j}$. We multiply A_{jx_t} to each element, which yields

$$\begin{aligned}\boldsymbol{\alpha}_t &= \begin{bmatrix} A_{1x_t}\boldsymbol{\alpha}_{t-1}^T\mathbf{B}_{:,1} \\ \vdots \\ A_{Jx_t}\boldsymbol{\alpha}_{t-1}^T\mathbf{B}_{:,J}\end{bmatrix}\\ &= \mathbf{A}_{:,x_t} \odot \begin{bmatrix} \boldsymbol{\alpha}_{t-1}^T\mathbf{B}_{:,1} \\ \vdots \\ \boldsymbol{\alpha}_{t-1}^T\mathbf{B}_{:,J}\end{bmatrix}.\end{aligned}$$

$\mathbf{B}_{:,1}, \cdots, \mathbf{B}_{:,J}$ as *columns* is unnatural (we are aligning different columns vertically), which motivates us to transpose this to obtain

$$\begin{aligned}&= \mathbf{A}_{:,x_t} \odot \begin{bmatrix}\boldsymbol{\alpha}_{t-1}^T\mathbf{B}_{:,1} & \cdots & \boldsymbol{\alpha}_{t-1}^T\mathbf{B}_{:,J}\end{bmatrix}^T\\ &= \mathbf{A}_{:,x_t} \odot \left(\boldsymbol{\alpha}_{t-1}^T\begin{bmatrix}\mathbf{B}_{:,1} & \cdots & \mathbf{B}_{:,J}\end{bmatrix}\right)^T\\ &= \mathbf{A}_{:,x_t} \odot \left(\boldsymbol{\alpha}_{t-1}^T\mathbf{B}\right)^T\\ &= \mathbf{A}_{:,x_t} \odot \mathbf{B}^T\boldsymbol{\alpha}_{t-1}.\end{aligned}$$

Alternatively, we can try a different way of visualizing $\sum_{k=1}^{J} \alpha_{t-1}(k) B_{kj}$:

$$\begin{bmatrix} & \vdots & \\ B_{j1}^T & \cdots & B_{jJ}^T \\ & \vdots & \end{bmatrix} \begin{bmatrix} \alpha_{t-1}(1) \\ \vdots \\ \alpha_{t-1}(J) \end{bmatrix}$$

or equivalently $\mathbf{B}_{j,:}^T \boldsymbol{\alpha}_{t-1}$. $\mathbf{B}_{j,:}^T$ should be interpreted as a *row vector* which corresponds to the j-th row of $\mathbf{B}^T$. This is slightly different from the notation in Section 3. Multiplying A_{jx_t} to each element again gives

$$\begin{aligned} \boldsymbol{\alpha}_t &= \begin{bmatrix} A_{1x_t} \mathbf{B}_{1,:}^T \boldsymbol{\alpha}_{t-1} \\ \vdots \\ A_{Jx_t} \mathbf{B}_{J,:}^T \boldsymbol{\alpha}_{t-1} \end{bmatrix} \\ &= \mathbf{A}_{:,x_t} \odot \begin{bmatrix} \mathbf{B}_{1,:}^T \boldsymbol{\alpha}_{t-1} \\ \vdots \\ \mathbf{B}_{J,:}^T \boldsymbol{\alpha}_{t-1} \end{bmatrix} \\ &= \mathbf{A}_{:,x_t} \odot \begin{bmatrix} \mathbf{B}_{1,:}^T \\ \vdots \\ \mathbf{B}_{J,:}^T \end{bmatrix} \boldsymbol{\alpha}_{t-1} \\ &= \mathbf{A}_{:,x_t} \odot \mathbf{B}^T \boldsymbol{\alpha}_{t-1}. \end{aligned}$$

Now we derive the vector form of $\beta_t(j) = \sum_{k=1}^{J} A_{kx_{t+1}} \beta_{t+1}(k) B_{jk}$. Recall that for $\mathbf{C} = \mathbf{AB}$, $C_{ij} = \sum_k A_{ik} B_{kj}$. Focus on the *colors*, not the letters! The color of the variable being summed over is olive, the variable before the olive variable is in red, and the one after it is in blue. To follow this, we reorder $\sum_{k=1}^{J} A_{kx_{t+1}} \beta_{t+1}(k) B_{jk}$ as $\sum_{k=1}^{J} B_{jk} A_{kx_{t+1}} \beta_{t+1}(k)$. $A_{kx_{t+1}}$ is the k-th element of the x_{t+1}-th column of $\mathbf{A}$, and $\beta_{t+1}(k)$ is the k-th element of the column vector $\boldsymbol{\beta}_{t+1}$. $A_{kx_{t+1}} \beta_{t+1}(k)$ is therefore the k-th element of $\mathbf{A}_{:,x_{t+1}} \odot \boldsymbol{\beta}_{t+1}$. This means

$$\begin{aligned} \beta_t(j) &= \sum_{k=1}^{J} A_{kx_{t+1}} \beta_{t+1}(k) B_{jk} \\ &= \sum_{k=1}^{J} B_{jk} \left[\mathbf{A}_{:,x_{t+1}} \odot \boldsymbol{\beta}_{t+1} \right]_k, \end{aligned}$$

and therefore

$$\boldsymbol{\beta}_t = \mathbf{B} \left(\mathbf{A}_{:,x_{t+1}} \odot \boldsymbol{\beta}_{t+1} \right).$$

7. (a) There is only one free variable j, so this indicates this is the j-th element of a vector. Visualizing $\sum_{i=1}^{N} A_{ij} x_i y_i$, we have

$$\begin{bmatrix} & A_{1j} & \\ \cdots & \vdots & \cdots \\ & A_{nj} & \end{bmatrix} \begin{bmatrix} x_1 y_1 \\ \vdots \\ x_n y_n \end{bmatrix}$$

which is not valid matrix multiplication. To fix this, we transpose $\mathbf{A}$ and obtain

$$\begin{bmatrix} \vdots \\ A_{j1}^T & \cdots & A_{jn}^T \\ \vdots \end{bmatrix} \begin{bmatrix} x_1 y_1 \\ \vdots \\ x_n y_n \end{bmatrix}.$$

Therefore, $\mathbf{v} = \mathbf{A}^T(\mathbf{x} \odot \mathbf{y})$.

(b) $\sum_{i=1}^N A_{ij} x_i$ is the same as (a) but with y_i removed. We have $\mathbf{v} = \mathbf{A}^T\mathbf{x}$.

(c) $\sum_{j=1}^N A_{ij} x_j y_j$ is almost the same as (a). Swapping i and j, we have $\sum_{i=1}^N A_{ji} x_i y_i$. The only difference is that A_{ij} in (a) is now A_{ji}, which means $\mathbf{v} = \mathbf{A}(\mathbf{x} \odot \mathbf{y})$.

(d) $\sum_{j=1}^N \sum_{k=1}^N A_{ki} B_{kj} x_j$ has only one free variable, i. Therefore, this is the i-th element of vector $\mathbf{v}$. First observe that

$$\begin{aligned} \sum_{j=1}^N \sum_{k=1}^N A_{ki} B_{kj} x_j &= \sum_{k=1}^N \sum_{j=1}^N A_{ki} B_{kj} x_j \\ &= \sum_{k=1}^N A_{ki} \sum_{j=1}^N B_{kj} x_j. \end{aligned}$$

$\sum_{j=1}^N B_{kj} x_j$ is the k-th element of $\mathbf{Bx}$, which means

$$\begin{aligned} &= \sum_{k=1}^N A_{ki} (\mathbf{Bx})_k \\ &= \sum_{k=1}^N A_{ik}^T (\mathbf{Bx})_k \\ &= \mathbf{A}^T\mathbf{Bx}. \end{aligned}$$

(e) $\sum_{i=1}^N \mathbf{x}^T \mathbf{A}_{i,:} \mathbf{A}_{i,:}^T$ has no free variables, so this represents the multiplication result itself. $\mathbf{x}^T$ does not depend on i, which allows us to rearrange this to $\mathbf{x}^T \sum_{i=1}^N \mathbf{A}_{i,:} \mathbf{A}_{i,:}^T$. Now consider $\mathbf{A}_{i,:} \mathbf{A}_{i,:}^T$:

$$\mathbf{A}_{i,:} \mathbf{A}_{i,:}^T = \begin{bmatrix} A_{i1} \\ \vdots \\ A_{in} \end{bmatrix} \begin{bmatrix} A_{i1} & \cdots & A_{in} \end{bmatrix} = \begin{bmatrix} A_{i1}A_{i1} & \cdots & A_{i1}A_{in} \\ \vdots & \ddots & \vdots \\ A_{in}A_{i1} & \cdots & A_{in}A_{in}, \end{bmatrix}$$

which means the (j,k)-th element of $\mathbf{A}_{i,:} \mathbf{A}_{i,:}^T$ is $A_{ij}A_{ik} = A_{ji}^T A_{ik}$. It naturally follows that the (j,k)-th element of $\sum_{i=1}^N \mathbf{A}_{i,:} \mathbf{A}_{i,:}^T$ is then simply $\sum_{i=1}^N A_{ji}^T A_{ik}$, which is the (j,k)-th element of $\mathbf{A}^T\mathbf{A}$. Therefore, $\sum_{i=1}^N \mathbf{A}_{i,:} \mathbf{A}_{i,:}^T = \mathbf{A}^T\mathbf{A}$, and we finally conclude that $\sum_{i=1}^N \mathbf{x}^T \mathbf{A}_{i,:} \mathbf{A}_{i,:}^T = \mathbf{x}^T\mathbf{A}^T\mathbf{A}$.

Note that $\mathbf{A}_{i,:}^T$ can be interpreted either as the i-th row of $\mathbf{A}^T$ in column vector, or as the i-th row of $\mathbf{A}$ in column vector transposed to a row vector. $\mathbf{A}_{i,:} \mathbf{A}_{i,:}^T$ is only defined for the latter case.

(f) $\sum_{j=1}^{N} A_{ij}x_iy_j$ has i as the only free variable, so this is a vector. x_i does not depend on j, which gives $\sum_{j=1}^{N} A_{ij}x_iy_j = x_i \sum_{j=1}^{N} A_{ij}y_j$. $\sum_{j=1}^{N} A_{ij}y_j$ is the i-th element of $\mathbf{Ay}$, and x_i is multiplied to this. Therefore, $\mathbf{v} = \mathbf{x} \odot \mathbf{Ay}$.

(g) $\sum_{i=1}^{N} \mathbf{A}_{:,i}\mathbf{B}_{:,i}^T\mathbf{x}$ does not have any free variables, so this is the result itself. $\mathbf{x}$ does not depend on i, so $\sum_{i=1}^{N} \mathbf{A}_{:,i}\mathbf{B}_{:,i}^T\mathbf{x} = \left(\sum_{i=1}^{N} \mathbf{A}_{:,i}\mathbf{B}_{:,i}^T\right)\mathbf{x}$. $\mathbf{A}_{:,i}\mathbf{B}_{:,i}^T$ can be seen as

$$\mathbf{A}_{:,i}\mathbf{B}_{:,i}^T = \begin{bmatrix} A_{1i} \\ \vdots \\ A_{ni} \end{bmatrix} \begin{bmatrix} B_{1i} & \cdots & B_{ni} \end{bmatrix} = \begin{bmatrix} A_{1i}B_{1i} & \cdots & A_{1i}B_{ni} \\ \vdots & \ddots & \vdots \\ A_{ni}B_{1i} & \cdots & A_{ni}B_{ni} \end{bmatrix}.$$

The (j,k)-th element of $\mathbf{A}_{:,i}\mathbf{B}_{:,i}^T$ is therefore $A_{ji}B_{ki} = A_{ji}B_{ik}^T$. Following the logic in (e), the answer is $\mathbf{AB}^T\mathbf{x}$.

(h) $x_iy_jA_{ij}$ has two free variables i and j. x_iy_j is the (i,j)-th element of $\mathbf{xy}^T$, and A_{ij} is the (i,j)-th element of $\mathbf{A}$. Therefore, the matrix whose (i,j)-th element is $x_iy_jA_{ij}$ is $\mathbf{xy}^T \odot \mathbf{A}$.

Note that there is another solution which has $x_iy_jA_{ij}$ as the (j,i)-th element, $\mathbf{yx}^T \odot \mathbf{A}^T$.

References

[1] C. A. Felippa. *Introduction to Finite Element Methods*. University of Colorado, 2004.

[2] I. Goodfellow, Y. Bengio, and A. Courville. *Deep Learning*. MIT Press, 2016.

[3] Z. Kolter and C. Do. Linear Algebra Review and Reference, Sep 2015.

[4] E. Learned-Miller. Vector, Matrix, and Tensor Derivatives, Mar 2017.

[5] T. Parr and J. Howard. The Matrix Calculus You Need For Deep Learning, 2018.

[6] K. B. Petersen and M. S. Pedersen. The Matrix Cookbook, Nov 2012.

[7] G. B. Thomas, M. D. Weir, and J. R. Hass. *Thomas' Calculus: Early Transcendentals*. Pearson Education, 13 edition, 2014.

[8] Wikipedia. Matrix calculus, Nov 2021.

[9] 김홍종. *미적분학 2+*. 서울대학교출판문화원, 2 edition, 2016.