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).

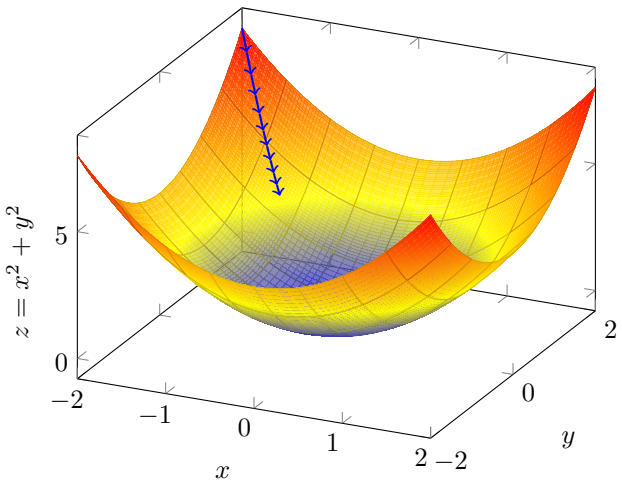

Figure 1: Climbing down $z = x^2 + y^2$ from point $(-2, 2, 8)$ following the *opposite* direction of the gradient vector.

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.

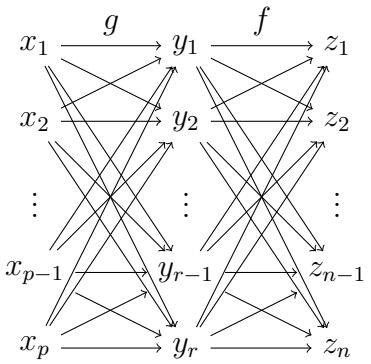

Figure 3: $\mathbf{z} = f(g(\mathbf{x}))$ visualized, where $\mathbf{z} = f(\mathbf{y})$ and $\mathbf{y} = g(\mathbf{x})$.

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:

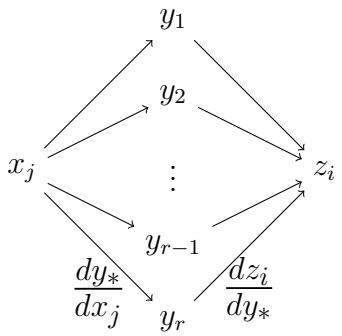

Figure 4: Chain rule visualized only considering z_i and x_j. y_* denotes any of $y_1, \cdots, y_r$.

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.

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

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