**Written originally for COMPSCI 181, Harvard University by Skyler Wu '24.*
Also intended for the brothers and sisters of the Stanford Statistics First-Year Cohort.

1 The Differential: A More Advanced Method for Matrix Derivatives

Up until now, we have relied on the element-wise approach to prove our matrix derivative identities. While the element-wise approach is undoubtedly intuitive, it may not be the best option for more complicated matrix derivatives. Thus, we now introduce a more advanced approach to matrix differentiation – employing the trace and the differential, and then extracting out the derivative. But first, we have to build up the theoretical tools.

1.1 Trace

Definition 1.1 (Trace)**.** For a $n \times n$ square matrix $\mathbf{A}$ with entries $[a_{ij}]$, the *trace* of $\mathbf{A}$ is defined as the sum of the entries on its main diagonal (i.e. from top left to bottom right). Mathematically, we have the following expression for trace [10, p. 11]:

$$\text{tr}(\mathbf{A}) = \sum_{i=1}^{n} a_{ii}$$

Theorem 1. If f is a scalar function, then $f = \text{tr}(f)$.

Proof. This is valid because if f is a scalar, then we can treat f as a 1×1 matrix. The sum of the entries of a 1×1 matrix (i.e., a scalar) is simply just the scalar itself. This fact will become very useful soon. □

The following properties will prove very useful for when we later calculate derivatives.

Properties of Trace [10, p. 11]:

1. $\text{tr}(\mathbf{A}) + \text{tr}(\mathbf{B}) = \text{tr}(\mathbf{A} + \mathbf{B})$, for $n \times n$ square matrices $\mathbf{A}$ and $\mathbf{B}$
2. $c\text{tr}(\mathbf{A}) = \text{tr}(c\mathbf{A})$, for $n \times n$ square matrix $\mathbf{A}$ and scalar c.
3. $\text{tr}(\mathbf{A}) = \text{tr}(\mathbf{A}^T)$, for $n \times n$ square matrix $\mathbf{A}$
4. $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$, so long as the dimensions of $\mathbf{A}$ and $\mathbf{B}$ are compatible.

The first two properties can be relatively directly extracted from the properties of matrix addition and scalar multiplication. The third comes from the fact that the diagonal entries of a square matrix and its transpose are the same. The fourth is a bit more involved. Note that Property 4 holds even if $\mathbf{AB}$ and $\mathbf{BA}$ have different dimensions, as is usually the case. Let us prove Property 4.

Theorem 2. $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$, so long as the dimensions of $\mathbf{A}$ and $\mathbf{B}$ are compatible [8, p. 5]. Note that this theorem holds for column or row vectors, provided that the dimensions are compatible.

Proof. Let $\mathbf{A} \in \mathbb{R}^{m \times n}$ and $\mathbf{B} \in \mathbb{R}^{n \times m}$. This is the only way for both $\mathbf{AB}$ and $\mathbf{BA}$ to be square matrices (for the trace operator to work), albeit with possibly different dimensions. To help us visualize the following derivation, let us draw out the shapes of $\mathbf{A}$ and $\mathbf{B}$.

$$\mathbf{A} = \begin{bmatrix} a_{11} & a_{12} & \cdots & a_{1n} \\ a_{21} & a_{22} & \cdots & a_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ a_{m1} & a_{m2} & \cdots & a_{mn} \end{bmatrix}, \quad \mathbf{B} = \begin{bmatrix} b_{11} & a_{12} & \cdots & a_{1m} \\ b_{21} & a_{22} & \cdots & a_{2m} \\ \vdots & \vdots & \ddots & \vdots \\ b_{n1} & a_{n2} & \cdots & a_{nm} \end{bmatrix}.$$

Let ab_{ii} represent an arbitrary diagonal entry of the resultant $m \times m$ square matrix $\mathbf{AB}$. By property of matrix multiplication, we have:

$$ab_{ii} = \sum_{k=1}^{n} a_{ik} b_{ki}.$$

It follows, from the definition of trace, that

$$\text{tr}(\mathbf{AB}) = \sum_{i=1}^{m} ab_{ii} = \sum_{i=1}^{m} \sum_{k=1}^{n} a_{ik} b_{ki}.$$

Now, let us unpack tr($\mathbf{BA}$). Let ba_{ii} represent an arbitrary diagonal entry of the resultant $n \times n$ square matrix $\mathbf{BA}$. By property of matrix multiplication, we have:

$$ba_{ii} = \sum_{k=1}^{m} b_{ik} a_{ki}$$

By definition of trace, we have:

$$\text{tr}(\mathbf{BA}) = \sum_{i=1}^{n} ba_{ii} = \sum_{i=1}^{n} \sum_{k=1}^{m} b_{ik} a_{ki}.$$

At this point, we are getting very close to the end. Let us rename our indexing variable i to k, and our indexing variable k to i. Then, we have:

$$\text{tr}(\mathbf{BA}) = \sum_{k=1}^{n} \sum_{i=1}^{m} b_{ki} a_{ik} = \sum_{i=1}^{m} \sum_{k=1}^{n} a_{ik} b_{ki}.$$

Now, it is clear that

$$\text{tr}(\mathbf{AB}) = \sum_{i=1}^{m} \sum_{k=1}^{n} a_{ik} b_{ki} = \text{tr}(\mathbf{BA})$$

□

The trace operator and its properties will soon become very useful for calculating complicated matrix derivatives. But first, we have to introduce the matrix differential.

1.2 The Differential

In AP Calculus, or another previous calculus course, the reader may have been introduced to the concept of the "differential." For example, in the scalar context of u-substitution, if we set $u = cos(x)$, then

$$du = -sin(x)dx.$$

In this example, du is the differential of u, while dx is the differential of x. The expression $du = -sin(x)dx$ tells us that if we increase x by an infinitesimally small dx, then u would increase by $-sin(x)dx$. Analagously, we proceed to *define* the differential in a matrix context.

Definition 1.2 (Matrix Differential)**.** Let $\mathbf{X}$ be an $m \times n$ matrix. The *matrix differential* of $\mathbf{X}$ is defined as follows. Intuitively, we are just taking the differential of each element in $\mathbf{X}$: [8, p. 7]

$$d\mathbf{X} = \begin{bmatrix} dx_{11} & dx_{12} & \cdots & dx_{1n} \\ dx_{21} & dx_{22} & \cdots & dx_{2n} \\ \vdots & \vdots & \ddots & \vdots \\ dx_{m1} & dx_{m2} & \cdots & dx_{mn} \end{bmatrix}$$

Foreshadowing later interactions between the matrix differential and the trace operator, we provide the following theorem. This theorem allows to move the trace operator in-or-out of the differential operator:

Theorem 3. [8, p. 7]

$$d\text{tr}(\mathbf{X}) = \text{tr}(d\mathbf{X}), \text{ for } m \times m \text{ square matrix } \mathbf{X}.$$

Proof. On the left-hand side of the equation, we have $d\text{tr}(\mathbf{X})$. From our understanding of trace, we know:

$$d\text{tr}(\mathbf{X}) = d(\sum_{i=1}^{m} x_{ii}) = dx_{11} + dx_{22} + \cdots + dx_{mm}.$$

On the right-hand side of the equation, we have $\text{tr}(d\mathbf{X})$. Per our definition of matrix differential, we know that the matrix differential of $\mathbf{X}$ is itself a matrix with the same dimensions as $\mathbf{X}$ that we can perform the trace operator on. Thus, we have:

$$\text{tr}(d\mathbf{X}) = \sum_{i=1}^{m} dx_{ii}.$$

It follows that $d\text{tr}(\mathbf{X}) = \text{tr}(d\mathbf{X})$. This theorem will be an essential step for many of our matrix derivatives. □

1.3 Extracting out the Derivative

Until now, we have not yet formally introduced any functions that we would like to take matrix derivatives of, yet. For the rest of this section, we will focus on scalar functions f that can take in either a matrix or a vector (which we will consider as an $m \times 1$ matrix) as input. For simplicity of notation, we will use $\mathbf{X}$ to defaultly represent a matrix, but any results we prove in this section for matrix $\mathbf{X}$ can directly apply, without modification, for vector $\mathbf{X}$, too.

Recall from an earlier math course the concept of a "scalar differential" of a function. If f is a scalar function of, say, three variables x, y, z (i.e., $f(x, y, z)$), then the *differential* of f is defined as the following: [5, p. 1]

$$df = \frac{\partial f}{\partial x}dx + \frac{\partial f}{\partial y}dy + \frac{\partial f}{\partial z}dz.$$

Naturally, this construal of the differential, or "exterior derivative", can also be extended to a scalar function f that takes in a $m \times n$ matrix $\mathbf{X}$ as input. Intuitively, we can think of the matrix $\mathbf{X}$ as just an entity of storing all of our function inputs. Then, we have, iterating through each entry in $\mathbf{X}$:

$$df = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\partial f}{\partial x_{ij}}dx_{ij}.$$

Working with differentials is oftentimes easier than directly working with derivatives. We provide the following properties of differentials without proof, for sake of space, and because they are the direct analogs of their corresponding derivative "rules" (i.e., constant, sum, product, etc.) These properties will come in very handy when we start solving matrix derivatives.

Properties of Matrix Differentials [10, p. 164]: For matrix functions/variables $\mathbf{X}$ and $\mathbf{Y}$, constant matrix $\mathbf{A}$, and scalar c, we have the following properties. Note that $\mathbf{X}$, $\mathbf{Y}$, and $\mathbf{A}$ can also be vectors, so long as the dimensions work out, in context.

1. $d\mathbf{A} = 0$, the analog of the derivative of a constant being equal to 0.
2. $d(c\mathbf{X}) = cd\mathbf{X}$, the analog of the constant multiplication rule for derivatives.

3. $d(\mathbf{X}+\mathbf{Y}) = d\mathbf{X} + d\mathbf{Y}$, the analog of the sum rule for derivatives.

4. $d(\mathbf{XY}) = (d\mathbf{X})\mathbf{Y} + \mathbf{X}(d\mathbf{Y})$, the analog of the product rule for derivatives.

Recall from the earliest sections of this paper that the derivative of a scalar function f with respect to a $m \times n$ matrix $\mathbf{X}$ should return an $n \times m$ matrix, per the conventions of *numerator layout*. In other words, $\frac{\partial f}{\partial \mathbf{X}}$ is a $n \times m$ matrix.

The following theorem connects the world of differentials with the world of derivatives. Notice how the trace, which we extensively discussed earlier, makes a return:

Theorem 4. For a scalar function f that takes as input a $m \times n$ matrix $\mathbf{X}$, we have [8, p. 7]:

$$df = \operatorname{tr}\left((\frac{\partial f}{\partial \mathbf{X}})d\mathbf{X}\right).$$

Remark: recall that $\frac{\partial f}{\partial \mathbf{X}}$ is a $n \times m$ matrix, and that $\mathbf{X}$ is a $m \times n$ matrix. Thus, $(\frac{\partial f}{\partial \mathbf{X}})d\mathbf{X}$ is an $n \times n$ square matrix, and thus the trace operator is viable.

Proof. From our discussions earlier in the section, we know that for a scalar function f that takes in a $m \times n$ matrix $\mathbf{X}$ as input, the differential, df, is defined as:

$$df = \sum_{i=1}^{m}\sum_{j=1}^{n} \frac{\partial f}{\partial x_{ij}} dx_{ij}.$$

From our definitions of $\frac{\partial f}{\partial \mathbf{X}}$ (under numerator layout) and $d\mathbf{X}$ we know:

$$\frac{\partial f}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial f}{\partial x_{11}} & \cdots & \frac{\partial f}{\partial x_{m1}} \\ \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial x_{1n}} & \cdots & \frac{\partial f}{\partial x_{mn}} \end{bmatrix}, \quad d\mathbf{X} = \begin{bmatrix} dx_{11} & \cdots & dx_{1n} \\ \vdots & \ddots & \vdots \\ dx_{m1} & \cdots & dx_{mn} \end{bmatrix}.$$

Consider the i^{th} diagonal entry of $(\frac{\partial f}{\partial \mathbf{X}})d\mathbf{X}$, which we will denote as z_i. By definition of matrix multiplication, we have the following:

$$z_i = \sum_{j=1}^{m} \frac{\partial f}{\partial x_{ji}} dx_{ji}.$$

By definition of trace, we have:

$$\operatorname{tr}\left((\frac{\partial f}{\partial \mathbf{X}})d\mathbf{X}\right) = \sum_{i=1}^{n} z_i = \sum_{i=1}^{n}\sum_{j=1}^{m} \frac{\partial f}{\partial x_{ji}} dx_{ji}.$$

To stay consistent with indices labeling, let us relabel i with j, and j with i. Then, we have:

$$\operatorname{tr}\left((\frac{\partial f}{\partial \mathbf{X}})d\mathbf{X}\right) = \sum_{j=1}^{n}\sum_{i=1}^{m} \frac{\partial f}{\partial x_{ij}} dx_{ij}.$$

Finally, we have proven our theorem:

$$df = \operatorname{tr}\left((\frac{\partial f}{\partial \mathbf{X}})d\mathbf{X}\right).$$

□

This theorem is the most important theorem in this entire section! It tells us that if we know the differential of a function f, df, and can express the differential as the trace of the expression for a matrix that ends with $d\mathbf{X}$, then we can directly extract out the derivative $\frac{\partial f}{\partial \mathbf{X}}$.

Everything we have discussed so far has been largely theoretical in nature. Now that we are equipped with all the theoretical concepts we need, we are ready for a practical derivative example.

1.4 Practical Example

Let us apply the properties of trace and differentials to the following computational example:

Theorem 5. $\dfrac{\partial \mathbf{a}^T\mathbf{X}\mathbf{a}}{\partial \mathbf{X}} = \mathbf{a}\mathbf{a}^T$, where $\mathbf{a} \in \mathbb{R}^{n\times 1}$ is a *constant* column vector, and $\mathbf{X} \in \mathbb{R}^{n\times n}$ is a square matrix [12, p. 10].

Proof. Let $f(\mathbf{X}) = \mathbf{a}^T\mathbf{X}\mathbf{a}$. Note that f is a scalar function.

By Theorem 1, the trace of a scalar function is just the function itself:

$$f = \text{tr}(f) = \text{tr}(\mathbf{a}^T\mathbf{X}\mathbf{a}).$$

By Theorem 3, we can move the differential operator inside of the trace operator:

$$df = d\left(\text{tr}(\mathbf{a}^T\mathbf{X}\mathbf{a})\right) = \text{tr}\left(d(\mathbf{a}^T\mathbf{X}\mathbf{a})\right).$$

By Theorem 2, we know that $\text{tr}(\mathbf{AB}) = \text{tr}(\mathbf{BA})$. Let us call $\mathbf{a}^T\mathbf{X}$ an alias variable $\mathbf{Q}$. Then, we have:

$$\text{tr}\left(d(\mathbf{Q}\mathbf{a})\right) = \text{tr}\left(d(\mathbf{a}\mathbf{Q})\right).$$

Substituting in the full expression for $\mathbf{Q}$, we have:

$$df = \text{tr}\left(d(\mathbf{a}\mathbf{a}^T\mathbf{X})\right).$$

**note:* even though $\mathbf{a}$ is a column vector, because of the dimensions of $\mathbf{a}$ and $\mathbf{X}$, the multiplication operation is still defined, so Theorem 2 still applies.

By the first property of Matrix Differentials, we know that for constant matrix $\mathbf{A}$, $d\mathbf{A} = 0$. By the fourth property of Matrix Differentials, we know that $d(\mathbf{XY}) = (d\mathbf{X})\mathbf{Y} + \mathbf{X}(d\mathbf{Y})$, for compatibly shaped matrix functions/variables $\mathbf{X}$ and $\mathbf{Y}$. We also know that $\mathbf{a}\mathbf{a}^T$ is a constant because $\mathbf{a}$ is a constant. Applying these principles to our problem, we have:

$$df = \text{tr}\left(d(\mathbf{a}\mathbf{a}^T)\mathbf{X} + (\mathbf{a}\mathbf{a}^T)d\mathbf{X}\right) = \text{tr}\left(0 + (\mathbf{a}\mathbf{a}^T)d\mathbf{X}\right) = \text{tr}\left(\mathbf{a}\mathbf{a}^T d\mathbf{X}\right).$$

Finally, by Theorem 4, we know that $df = \text{tr}\left((\frac{\partial f}{\partial \mathbf{X}})d\mathbf{X}\right)$. As such, looking at the expression for df from above, we can directly extract out our derivative:

$$\frac{\partial f}{\partial \mathbf{X}} = \mathbf{a}\mathbf{a}^T.$$

□

1.5 Remarks

The trace and differential-based tools that we have constructed in this section allow us to compute matrix derivatives much more efficiently than the element-wise approach. The trace and differential-based approach to matrix differentiation allows to tackle much more complicated derivatives, too.

References

[1] Abadir, Karim M., and Jan R. Magnus. *Matrix Algebra. Vol. 1.* Cambridge University Press, 2005.

[2] Adams, Ryan. "COS 302 Precept 6." COS 302 / SML 305: Mathematics for Numerical Computing and Machine Learning, Princeton University, 2020, www.cs.princeton.edu/courses/archive/spring20/cos302/files/COS_302_Precept_6.pdf.

[3] Ahmadi, Amir Ali. "Lecture 4." ORF523: Convex and Conic Optimization. Princeton University, February 16, 2016. https://www.princeton.edu/ aaa/Public/Teaching/ORF523/S16/ORF523_S16_Lec4_gh.pdf.

[4] Barnes, Randal J. "Matrix differentiation." Springs Journal (2006): 1-9.

[5] Clelland, Jeanne N. "Lecture 1: Differential Forms." MATH 4470/5470: Intro to Partial Differential Equations. University of Colorado, Boulder, n.d. https://math.colorado.edu/ jnc/lecture1.pdf.

[6] Deuschle, William J. 2019. *Undergraduate Fundamentals of Machine Learning.* Bachelor's thesis, Harvard College (CS181 course textbook).

[7] Green, Larry. "Cofactors." Matrices and Applications, Lake Tahoe Community College, ltcconline.net/greenl/courses/203/MatricesApps/cofactors.htm.

[8] Hu, Pili. *Matrix Calculus: Derivation and Simple Application.* Technical report, City University of Hong Kong, 2012.

[9] Lay, David C., Steven R. Lay, and Judi J. McDonald. *Linear Algebra and its Applications.* (2016).

[10] Magnus, Jan R., and Heinz Neudecker. *Matrix Differential Calculus with Applications in Statistics and Econometrics.* John Wiley & Sons, 2019.

[11] Minka, Thomas P. Old and New Matrix Algebra Useful for Statistics. https://tminka.github.io/papers/matrix/minka-matrix.pdf.

[12] Petersen, Kaare Brandt, and Michael Syskind Pedersen. *The Matrix Cookbook.* Technical University of Denmark, 2012.