Matrix Calculus - Notes on the Derivative of a Trace

Johannes Traa

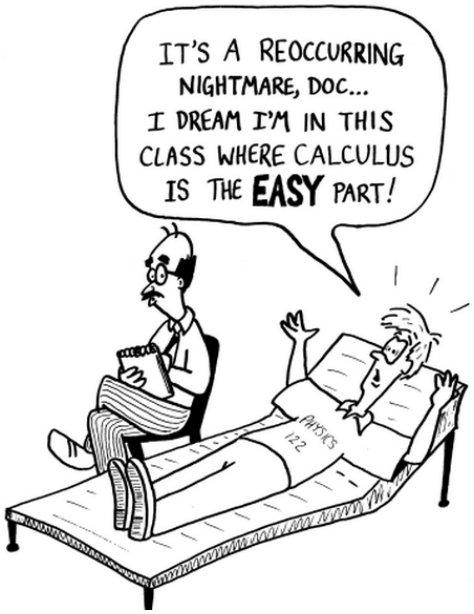
This write-up elucidates the rules of matrix calculus for expressions involving the trace of a function of a matrix $\mathbf{X}$:

$$f = tr\left[g(\mathbf{X})\right] \ . \tag{1}$$

We would like to take the derivative of f with respect to $\mathbf{X}$:

$$\frac{\partial f}{\partial \mathbf{X}} = ? \ . \tag{2}$$

One strategy is to write the trace expression as a scalar using index notation, take the derivative, and re-write in matrix form. An easier way is to reduce the problem to one or more smaller problems where the results for simpler derivatives can be applied. It's brute-force vs bottom-up.



MATRIX-VALUED DERIVATIVE

The derivative of a scalar f with respect to a matrix $\mathbf{X} \in \mathbb{R}^{M \times N}$ can be written as:

$$\frac{\partial f}{\partial \mathbf{X}} = \begin{bmatrix} \frac{\partial f}{\partial X_{11}} & \frac{\partial f}{\partial X_{12}} & \cdots & \frac{\partial f}{\partial X_{1N}} \\ \frac{\partial f}{\partial X_{21}} & \frac{\partial f}{\partial X_{22}} & \cdots & \frac{\partial f}{\partial X_{2N}} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial f}{\partial X_{M1}} & \frac{\partial f}{\partial X_{M2}} & \cdots & \frac{\partial f}{\partial X_{MN}} \end{bmatrix} \tag{3}$$

So, the result is the same size as **X**.

Matrix and Index Notation

It is useful to be able to convert between matrix notation and index notation. For example, the product **AB** has elements:

$$[\mathbf{AB}]_{ik} = \sum_j A_{ij} B_{jk} \ , \tag{4}$$

and the matrix product $\mathbf{ABC}^T$ has elements:

$$\left[\mathbf{ABC}^T\right]_{il} = \sum_j A_{ij} \left[\mathbf{BC}^T\right]_{jl} = \sum_j A_{ij} \sum_k B_{jk} C_{lk} = \sum_j \sum_k A_{ij} B_{jk} C_{lk} \ . \tag{5}$$

First-Order Derivatives

Example 1

Consider this example:

$$f = tr\left[\mathbf{AXB}\right] \ . \tag{6}$$

We can write this using index notation as:

$$f = \sum_i \left[\mathbf{AXB}\right]_{ii} = \sum_i \sum_j A_{ij} \left[\mathbf{XB}\right]_{ji} = \sum_i \sum_j A_{ij} \sum_k X_{jk} B_{ki} = \sum_i \sum_j \sum_k A_{ij} X_{jk} B_{ki} \ . \tag{7}$$

Taking the derivative with respect to X_{jk}, we get:

$$\frac{\partial f}{\partial X_{jk}} = \sum_i A_{ij} B_{ki} = \left[\mathbf{BA}\right]_{kj} \ . \tag{8}$$

The result has to be the same size as **X**, so we know that the indices of the rows and columns must be j and k, respectively. This means we have to transpose the result above to write the derivative in matrix form as:

$$\frac{\partial tr\left[\mathbf{AXB}\right]}{\partial \mathbf{X}} = \mathbf{A}^T \mathbf{B}^T \ . \tag{9}$$

EXAMPLE 2

Similarly, we have:

$$f = tr\left[\mathbf{A}\mathbf{X}^T\mathbf{B}\right] = \sum_i \sum_j \sum_k A_{ij} X_{kj} B_{ki} \ , \tag{10}$$

so that the derivative is:

$$\frac{\partial f}{\partial X_{kj}} = \sum_i A_{ij} B_{ki} = [\mathbf{B}\mathbf{A}]_{kj} \ , \tag{11}$$

The $\mathbf{X}$ term appears in (10) with indices kj, so we need to write the derivative in matrix form such that k is the row index and j is the column index. Thus, we have:

$$\frac{\partial\, tr\left[\mathbf{A}\mathbf{X}^T\mathbf{B}\right]}{\partial \mathbf{X}} = \mathbf{B}\mathbf{A} \ . \tag{12}$$

MULTIPLE-ORDER

Now consider a more complicated example:

$$f = tr\left[\mathbf{A}\mathbf{X}\mathbf{B}\mathbf{X}\mathbf{C}^T\right] \tag{13}$$

$$= \sum_i \sum_j \sum_k \sum_l \sum_m A_{ij} X_{jk} B_{kl} X_{lm} C_{im} \ . \tag{14}$$

The derivative has contributions from both appearances of $\mathbf{X}$.

TAKE 1

In index notation:

$$\frac{\partial f}{\partial X_{jk}} = \sum_i \sum_l \sum_m A_{ij} B_{kl} X_{lm} C_{im} = \left[\mathbf{B}\mathbf{X}\mathbf{C}^T\mathbf{A}\right]_{kj} \ , \tag{15}$$

$$\frac{\partial f}{\partial X_{lm}} = \sum_i \sum_j \sum_k A_{ij} X_{jk} B_{kl} C_{im} = \left[\mathbf{C}^T\mathbf{A}\mathbf{X}\mathbf{B}\right]_{ml} \ . \tag{16}$$

Transposing appropriately and summing the terms together, we have:

$$\frac{\partial\, tr\left[\mathbf{A}\mathbf{X}\mathbf{B}\mathbf{X}\mathbf{C}^T\right]}{\partial \mathbf{X}} = \mathbf{A}^T\mathbf{C}\mathbf{X}^T\mathbf{B}^T + \mathbf{B}^T\mathbf{X}^T\mathbf{A}^T\mathbf{C} \ . \tag{17}$$

TAKE 2

We can skip this tedious process by applying (9) for each appearance of $\mathbf{X}$:

$$\frac{\partial\, tr\left[\mathbf{A}\mathbf{X}\mathbf{B}\mathbf{X}\mathbf{C}^T\right]}{\partial \mathbf{X}} = \frac{\partial\, tr\left[\mathbf{A}\mathbf{X}\mathbf{D}\right]}{\partial \mathbf{X}} + \frac{\partial\, tr\left[\mathbf{E}\mathbf{X}\mathbf{C}^T\right]}{\partial \mathbf{X}} = \mathbf{A}^T\mathbf{D}^T + \mathbf{E}^T\mathbf{C} \ . \tag{18}$$

where $\mathbf{D} = \mathbf{B}\mathbf{X}\mathbf{C}^T$ and $\mathbf{E} = \mathbf{A}\mathbf{X}\mathbf{B}$. So we just evaluate the matrix derivative for each appearance of $\mathbf{X}$ assuming that everything else is a constant (including other $\mathbf{X}$'s). To see why this rule is useful, consider the following beast:

$$f = tr\left[\mathbf{A}\mathbf{X}\mathbf{X}^T\mathbf{B}\mathbf{C}\mathbf{X}^T\mathbf{X}\mathbf{C}\right] \ . \tag{19}$$

We can immediately write down the derivative using (9) and (12):

$$\frac{\partial\, tr\left[\mathbf{A}\mathbf{X}\mathbf{X}^T\mathbf{B}\mathbf{C}\mathbf{X}^T\mathbf{X}\mathbf{C}\right]}{\partial \mathbf{X}} = (\mathbf{A})^T\left(\mathbf{X}^T\mathbf{B}\mathbf{C}\mathbf{X}^T\mathbf{X}\mathbf{C}\right)^T + \left(\mathbf{B}\mathbf{C}\mathbf{X}^T\mathbf{X}\mathbf{C}\right)(\mathbf{A}\mathbf{X}) + (\mathbf{X}\mathbf{C})\left(\mathbf{A}\mathbf{X}\mathbf{X}^T\mathbf{B}\mathbf{C}\right) + \left(\mathbf{A}\mathbf{X}\mathbf{X}^T\mathbf{B}\mathbf{C}\mathbf{X}^T\right)^T(\mathbf{C})^T \tag{20}$$

$$= \mathbf{A}\mathbf{C}^T\mathbf{X}^T\mathbf{X}\mathbf{C}^T\mathbf{B}^T\mathbf{X} + \mathbf{B}\mathbf{C}\mathbf{X}^T\mathbf{X}\mathbf{C}\mathbf{A}\mathbf{X} + \mathbf{X}\mathbf{C}\mathbf{A}\mathbf{X}\mathbf{X}^T\mathbf{B}\mathbf{C} + \mathbf{X}\mathbf{C}^T\mathbf{B}^T\mathbf{X}\mathbf{X}^T\mathbf{A}^T\mathbf{C}^T \ . \tag{21}$$

FROBENIUS NORM

The Frobenius norm shows up when we have an optimization problem involving a matrix factorization and we want to minimize a sum-of-squares error criterion:

$$f = \sum_i \sum_k \left(X_{ik} - \sum_j W_{ij}H_{jk}\right)^2 = \|\mathbf{X} - \mathbf{W}\mathbf{H}\|_F^2 = tr\left[(\mathbf{X} - \mathbf{W}\mathbf{H})(\mathbf{X} - \mathbf{W}\mathbf{H})^T\right] \ . \tag{22}$$

We can work with the expression in index notation, but it's easier to work directly with matrices and apply the results derived earlier. Suppose we want to find the derivative with respect to $\mathbf{W}$. Expanding the matrix outer product, we have:

$$f = tr\left[\mathbf{X}\mathbf{X}^T\right] - tr\left[\mathbf{X}\mathbf{H}^T\mathbf{W}^T\right] - tr\left[\mathbf{W}\mathbf{H}\mathbf{X}^T\right] + tr\left[\mathbf{W}\mathbf{H}\mathbf{H}^T\mathbf{W}^T\right] \ . \tag{23}$$

Applying (9) and (12), we easily deduce that:

$$\frac{\partial\, tr\left[(\mathbf{X} - \mathbf{W}\mathbf{H})(\mathbf{X} - \mathbf{W}\mathbf{H})^T\right]}{\partial \mathbf{W}} = -2\mathbf{X}\mathbf{H}^T + 2\mathbf{W}\mathbf{H}\mathbf{H}^T \ . \tag{24}$$

LOG

Consider this trace expression:

$$f = tr\left[\mathbf{V}^T \log(\mathbf{A}\mathbf{X}\mathbf{B})\right] = \sum_i \sum_j V_{ij} \log\left(\sum_m \sum_n A_{im}X_{mn}B_{nj}\right) \ . \tag{25}$$

Taking the derivative with respect to X_{mn}, we get:

$$\frac{\partial f}{\partial X_{mn}} = \sum_i \sum_j \left(\frac{V_{ij}}{\sum_m \sum_n A_{im}X_{mn}B_{nj}}\right) A_{im}B_{nj} = \sum_i \sum_j A_{im}\left(\frac{\mathbf{V}}{\mathbf{A}\mathbf{X}\mathbf{B}}\right)_{ij} B_{nj} \ . \tag{26}$$

Thus:

$$\frac{\partial\, tr\left[\mathbf{V}^T \log\left(\mathbf{AXB}\right)\right]}{\partial \mathbf{X}} = \mathbf{A}^T \left(\frac{\mathbf{V}}{\mathbf{AXB}}\right) \mathbf{B}^T \ . \tag{27}$$

Similarly:

$$\frac{\partial\, tr\left[\mathbf{V}^T \log\left(\mathbf{AX}^T\mathbf{B}\right)\right]}{\partial \mathbf{X}} = \mathbf{B} \left(\frac{\mathbf{V}}{\mathbf{AX}^T\mathbf{B}}\right)^T \mathbf{A} \ . \tag{28}$$

These bare a spooky resemblance to (9) and (12).

DIAG

EXAMPLE 1

Consider the tricky case of a $diag\left(-\right)$ operator:

$$f = tr\left[\mathbf{A}\, diag\left(\mathbf{x}\right) \mathbf{B}\right] = \sum_i \sum_j A_{ij} x_j B_{ji} \ . \tag{29}$$

Taking the derivative, we have:

$$\frac{\partial f}{\partial x_j} = \sum_i A_{ij} B_{ji} = \left[\left(\mathbf{A}^T \odot \mathbf{B}\right) \mathbf{1}\right]_j \ . \tag{30}$$

So we can write:

$$\frac{\partial\, tr\left[\mathbf{A}\, diag\left(\mathbf{x}\right) \mathbf{B}\right]}{\partial \mathbf{x}} = \left(\mathbf{A}^T \odot \mathbf{B}\right) \mathbf{1} \ . \tag{31}$$

EXAMPLE 2

Consider the following take on the last example:

$$f = tr\left[J\, \mathbf{A}\, diag\left(\mathbf{x}\right) \mathbf{B}\right] = \sum_i \sum_j \sum_k A_{ij} x_j B_{jk} \ , \tag{32}$$

where J is the matrix of ones.

TAKE 1

Taking the derivative, we have:

$$\frac{\partial f}{\partial x_j} = \sum_i \sum_k A_{ij} B_{jk} = \left(\sum_i A_{ij}\right)\left(\sum_k B_{jk}\right) = \left[\mathbf{A}^T \mathbf{1} \odot \mathbf{B1}\right]_j \ . \tag{33}$$

So we can write:

$$\frac{\partial\, tr\left[J\, \mathbf{A}\, diag\left(\mathbf{x}\right) \mathbf{B}\right]}{\partial \mathbf{x}} = \mathbf{A}^T \mathbf{1} \odot \mathbf{B1} \ . \tag{34}$$

TAKE 2

We could have derived this result from the previous example using the rotation property of the trace operator:

$$f = tr\left[J\,\mathbf{A}\,diag\left(\mathbf{x}\right)\mathbf{B}\right] = tr\left[\mathbf{1}\mathbf{1}^T\,\mathbf{A}\,diag\left(\mathbf{x}\right)\mathbf{B}\right] = tr\left[\mathbf{1}^T\mathbf{A}\,diag\left(\mathbf{x}\right)\mathbf{B}\mathbf{1}\right] = tr\left[\mathbf{a}^T\,diag\left(\mathbf{x}\right)\mathbf{b}\right]\ , \tag{35}$$

where we have defined $\mathbf{a} = \mathbf{A}^T\mathbf{1}$ and $\mathbf{b} = \mathbf{B}\,\mathbf{1}$. Applying (31), we have:

$$\frac{\partial\ tr\left[\mathbf{a}^T\,diag\left(\mathbf{x}\right)\mathbf{b}\right]}{\partial\,\mathbf{x}} = \left(\mathbf{a}\odot\mathbf{b}\right)1 = \mathbf{A}^T\mathbf{1}\odot\mathbf{B}\,\mathbf{1}\ . \tag{36}$$

EXAMPLE 3

Consider a more complicated example:

$$f = tr\left[\mathbf{V}^T\log\left(\mathbf{A}\,diag\left(\mathbf{x}\right)\mathbf{B}\right)\right] = \sum_i\sum_j V_{ij}\log\left(\sum_m A_{im}x_mB_{mj}\right)\ . \tag{37}$$

Taking the derivative with respect to x_m, we have:

$$\frac{\partial f}{\partial x_m} = \sum_i\sum_j\left(\frac{V_{ij}}{\sum\limits_m A_{im}x_mB_{mj}}\right)A_{im}B_{mj} \tag{38}$$

$$= \sum_j\left[\sum_i A_{im}\left(\frac{\mathbf{V}}{\mathbf{A}\,diag\left(\mathbf{x}\right)\mathbf{B}}\right)_{ij}\right]B_{mj} \tag{39}$$

$$= \left[\left(\mathbf{A}^T\left(\frac{\mathbf{V}}{\mathbf{A}\,diag\left(\mathbf{x}\right)\mathbf{B}}\right)\odot\mathbf{B}\right)\mathbf{1}\right]_m\ . \tag{40}$$

The final result is:

$$\frac{\partial\ tr\left[\mathbf{V}^T\log\left(\mathbf{A}\,diag\left(\mathbf{x}\right)\mathbf{B}\right)\right]}{\partial\,\mathbf{x}} = \left(\mathbf{A}^T\left(\frac{\mathbf{V}}{\mathbf{A}\,diag\left(\mathbf{x}\right)\mathbf{B}}\right)\odot\mathbf{B}\right)\mathbf{1}\ . \tag{41}$$

EXAMPLE 4

How about when we have a trace composed of a sum of expressions, each of which depends on what row of a matrix $\mathbf{B}$ is chosen:

$$f = tr\left[\sum_k\mathbf{V}^T\log\left(\mathbf{A}\,diag\left(\mathbf{B}_{k:}\,\mathbf{X}\right)\mathbf{C}\right)\right] = \sum_k\sum_i\sum_j V_{ij}\log\left(\sum_m A_{im}\left(\sum_n B_{kn}X_{nm}\right)C_{mj}\right)\ . \tag{42}$$

Taking the derivative, we get:

$$\frac{\partial f}{\partial X_{nm}} = \sum_k \sum_i \sum_j \left(\frac{V_{ij}}{\sum_m A_{im} \left(\sum_n B_{kn} X_{nm} \right) C_{mj}} \right) A_{im} B_{kn} C_{mj} \tag{43}$$

$$= \sum_k B_{kn} \sum_i \sum_j \left(\frac{\mathbf{V}}{\mathbf{A}\, diag\left(\mathbf{B}_{k:}\mathbf{X}\right)\mathbf{C}} \right)_{ij} A_{im} C_{mj} \tag{44}$$

$$= \sum_k B_{kn} \left[\mathbf{1}^T \left(\left(\frac{\mathbf{V}}{\mathbf{A}\, diag\left(\mathbf{B}_{k:}\mathbf{X}\right)\mathbf{C}} \right)^T \mathbf{A} \odot \mathbf{C}^T \right) \right]_{km} , \tag{45}$$

so we can write:

$$\frac{\partial\, tr\left[\sum_k \mathbf{V}^T \log\left(\mathbf{A}\, diag\left(\mathbf{B}_{k:}\mathbf{X}\right)\mathbf{C}\right) \right]}{\partial \mathbf{X}} = \mathbf{B}^T \begin{bmatrix} \mathbf{1}^T \left(\left(\frac{\mathbf{V}}{\mathbf{A}\, diag\left(\mathbf{B}_{1:}\mathbf{X}\right)\mathbf{C}} \right)^T \mathbf{A} \odot \mathbf{C}^T \right) \\ \vdots \\ \mathbf{1}^T \left(\left(\frac{\mathbf{V}}{\mathbf{A}\, diag\left(\mathbf{B}_{K:}\mathbf{X}\right)\mathbf{C}} \right)^T \mathbf{A} \odot \mathbf{C}^T \right) \end{bmatrix} . \tag{46}$$

CONCLUSION

We've learned two things. First, if we don't know how to find the derivative of an expression using matrix calculus directly, we can always fall back on index notation and convert back to matrices at the end. This reduces a potentially unintuitive matrix-valued problem into one involving scalars, which we are used to. Second, it's less painful to massage an expression into a familiar form and apply previously-derived identities.