Similarly, it can be shown that

$$\begin{vmatrix} A & B \\ C & D \end{vmatrix} = |D||A - BD^{-1}C| \tag{1.48}$$

These formulas are called product rules for determinants. They were first given by the Russian-born mathematician Issai Schur in a German paper [Sch17] that was reprinted in English in [Sch86].

1.1.3 Matrix calculus

In our first calculus course, we learned the mathematics of derivatives and integrals and how to apply those concepts to scalars. We can also apply the mathematics of calculus to vectors and matrices. Some aspects of matrix calculus are identical to scalar calculus, but some scalar calculus concepts need to be extended in order to derive formulas for matrix calculus.

As intuition would lead us to believe, the time derivative of a matrix is simply equal to the matrix of the time derivatives of the individual matrix elements. Also, the integral of a matrix is equal to the matrix of the integrals of the individual matrix elements. In other words, assuming that A is an $m \times n$ matrix, we have

$$\begin{aligned} \dot{A}(t) &= \begin{bmatrix} \dot{A}_{11}(t) & \cdots & \dot{A}_{1n}(t) \\ \vdots & \ddots & \vdots \\ \dot{A}_{n1}(t) & \cdots & \dot{A}_{nn}(t) \end{bmatrix} \\ \int A(t)\,dt &= \begin{bmatrix} \int A_{11}(t)\,dt & \cdots & \int A_{1n}(t)\,dt \\ \vdots & \ddots & \vdots \\ \int A_{n1}(t)\,dt & \cdots & \int A_{nn}(t)\,dt \end{bmatrix} \end{aligned} \tag{1.49}$$

Next we will compute the time derivative of the inverse of a matrix. Suppose that matrix $A(t)$, which we will denote as A, has elements that are functions of time. We know that $AA^{-1} = I$; that is, AA^{-1} is a constant matrix and therefore has a time derivative of zero. But the time derivative of AA^{-1} can be computed as

$$\frac{d}{dt}(AA^{-1}) = \dot{A}A^{-1} + A\frac{d}{dt}(A^{-1}) \tag{1.50}$$

Since this is zero, we can solve for $d(A^{-1})/dt$ as

$$\frac{d}{dt}(A^{-1}) = -A^{-1}\dot{A}A^{-1} \tag{1.51}$$

Note that for the special case of a scalar A, this reduces to the familiar equation

$$\begin{aligned} \frac{d}{dt}(1/A) &= \frac{\partial(1/A)}{\partial A}\frac{dA}{dt} \\ &= -\dot{A}/A^2 \end{aligned} \tag{1.52}$$

Now suppose that x is an $n \times 1$ vector and $f(x)$ is a scalar function of the elements of x. Then

$$\frac{\partial f}{\partial x} = \begin{bmatrix} \partial f/\partial x_1 & \cdots & \partial f/\partial x_n \end{bmatrix} \tag{1.53}$$

Even though x is a column vector, $\partial f/\partial x$ is a row vector. The converse is also true – if x is a row vector, then $\partial f/\partial x$ is a column vector. Note that some authors define this the other way around. That is, they say that if x is a column vector then $\partial f/\partial x$ is also a column vector. There is no accepted convention for the definition of the partial derivative of a scalar with respect to a vector. It does not really matter which definition we use as long as we are consistent. In this book, we will use the convention described by Equation (1.53).

Now suppose that A is an $m \times n$ matrix and $f(A)$ is a scalar. Then the partial derivative of a scalar with respect to a matrix can be computed as follows:

$$\frac{\partial f}{\partial A} = \begin{bmatrix} \partial f/\partial A_{11} & \cdots & \partial f/\partial A_{1n} \\ \vdots & \ddots & \vdots \\ \partial f/\partial A_{m1} & \cdots & \partial f/\partial A_{mn} \end{bmatrix} \tag{1.54}$$

With these definitions we can compute the partial derivative of the dot product of two vectors. Suppose x and y are n-element column vectors. Then

$$\begin{aligned} x^T y &= x_1 y_1 + \cdots + x_n y_n \\ \frac{\partial(x^T y)}{\partial x} &= \left[\begin{array}{ccc} \partial(x^T y)/\partial x_1 & \cdots & \partial(x^T y)/\partial x_n \end{array} \right] \\ &= \left[\begin{array}{ccc} y_1 & \cdots & y_n \end{array} \right] \\ &= y^T \end{aligned} \tag{1.55}$$

Likewise, we can obtain

$$\frac{\partial(x^T y)}{\partial y} = x^T \tag{1.56}$$

Now we will compute the partial derivative of a quadratic with respect to a vector. First write the quadratic as follows:

$$\begin{aligned} x^T A x &= \left[\begin{array}{ccc} x_1 & \cdots & x_n \end{array} \right] \begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & \ddots & \vdots \\ A_{n1} & \cdots & A_{nn} \end{bmatrix} \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \\ &= \left[\begin{array}{ccc} \sum_i x_i A_{i1} & \cdots & \sum_i x_i A_{in} \end{array} \right] \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix} \\ &= \sum_{i,j} x_i x_j A_{ij} \end{aligned} \tag{1.57}$$

Now take the partial derivative of the quadratic as follows:

$$\begin{aligned} \frac{\partial(x^T A x)}{\partial x} &= \left[\begin{array}{ccc} \partial(x^T A x)/\partial x_1 & \cdots & \partial(x^T A x)/\partial x_n \end{array} \right] \\ &= \left[\begin{array}{ccc} \sum_j x_j A_{1j} + \sum_i x_i A_{i1} & \cdots & \sum_j x_j A_{nj} + \sum_i x_i A_{in} \end{array} \right] \\ &= \left[\begin{array}{ccc} \sum_j x_j A_{1j} & \cdots & \sum_j x_j A_{nj} \end{array} \right] + \left[\begin{array}{ccc} \sum_i x_i A_{i1} & \cdots & \sum_i x_i A_{in} \end{array} \right] \\ &= x^T A^T + x^T A \end{aligned} \tag{1.58}$$

If A is symmetric, as it often is in quadratic expressions, then $A = A^T$ and the above expression simplifies to

$$\frac{\partial(x^T A x)}{\partial x} = 2x^T A \qquad \text{if } A = A^T \tag{1.59}$$

Next we define the partial derivative of a vector with respect to another vector. Suppose $g(x) = \begin{bmatrix} g_1(x) \\ \vdots \\ g_m(x) \end{bmatrix}$ and $x = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}$. Then

$$\frac{\partial g}{\partial x} = \begin{bmatrix} \partial g_1/\partial x_1 & \cdots & \partial g_1/\partial x_n \\ \vdots & & \vdots \\ \partial g_m/\partial x_1 & \cdots & \partial g_m \partial x_n \end{bmatrix} \tag{1.60}$$

If either $g(x)$ or x is transposed, then the partial derivative is also transposed.

$$\begin{aligned} \frac{\partial g^T}{\partial x} &= \left(\frac{\partial g}{\partial x}\right)^T \\ \frac{\partial g}{\partial x^T} &= \left(\frac{\partial g}{\partial x}\right)^T \\ \frac{\partial g^T}{\partial x^T} &= \frac{\partial g}{\partial x} \end{aligned} \tag{1.61}$$

With these definitions, the following important equalities can be derived. Suppose A is an $m \times n$ matrix and x is an $n \times 1$ vector. Then

$$\begin{aligned} \frac{\partial(Ax)}{\partial x} &= A \\ \frac{\partial(x^T A)}{\partial x} &= A \end{aligned} \tag{1.62}$$

Now we suppose that A is an $m \times n$ matrix, B is an $n \times n$ matrix, and we want to compute the partial derivative of $\text{Tr}(ABA^T)$ with respect to A. First compute ABA^T as follows:

$$\begin{aligned} ABA^T &= \begin{bmatrix} A_{11} & \cdots & A_{1n} \\ \vdots & & \vdots \\ A_{m1} & \cdots & A_{mn} \end{bmatrix} \begin{bmatrix} B_{11} & \cdots & B_{1n} \\ \vdots & \ddots & \vdots \\ B_{n1} & \cdots & B_{nn} \end{bmatrix} \begin{bmatrix} A_{11} & \cdots & A_{m1} \\ \vdots & \ddots & \vdots \\ A_{1n} & \cdots & A_{nn} \end{bmatrix} \\ &= \begin{bmatrix} \sum_{j,k} A_{1k} B_{kj} A_{1j} & \cdots & \sum_{j,k} A_{1k} B_{kj} A_{mj} \\ \vdots & & \vdots \\ \sum_{j,k} A_{mk} B_{kj} A_{1j} & \cdots & \sum_{j,k} A_{mk} B_{kj} A_{mj} \end{bmatrix} \end{aligned} \tag{1.63}$$

From this we see that the trace of ABA^T is given as

$$\text{Tr}(ABA^T) = \sum_{i,j,k} A_{ik} B_{kj} A_{ij} \tag{1.64}$$

Its partial derivative with respect to A can be computed as

$$\begin{aligned}
\frac{\partial \mathrm{Tr}(ABA^T)}{\partial A} &= \begin{bmatrix} \partial \mathrm{Tr}(ABA^T)/\partial A_{11} & \cdots & \partial \mathrm{Tr}(ABA^T)/\partial A_{1n} \\ \vdots & & \vdots \\ \partial \mathrm{Tr}(ABA^T)/\partial A_{m1} & \cdots & \partial \mathrm{Tr}(ABA^T)/\partial A_{mn} \end{bmatrix} \\
&= \begin{bmatrix} \sum_j A_{1j}B_{1j} + \sum_k A_{1k}B_{k1} & \cdots & \sum_j A_{1j}B_{nj} + \sum_k A_{1k}B_{kn} \\ \vdots & \ddots & \vdots \\ \sum_j A_{mj}B_{1j} + \sum_k A_{mk}B_{k1} & \cdots & \sum_j A_{mj}B_{nj} + \sum_k A_{mk}B_{kn} \end{bmatrix} \\
&= \begin{bmatrix} \sum_j A_{1j}B_{1j} & \cdots & \sum_j A_{1j}B_{nj} \\ \vdots & & \vdots \\ \sum_j A_{mj}B_{1j} & \cdots & \sum_j A_{mj}B_{nj} \end{bmatrix} + \\
&\quad \begin{bmatrix} \sum_k A_{1k}B_{k1} & \cdots & \sum_k A_{1k}B_{kn} \\ \vdots & & \vdots \\ \sum_k A_{mk}B_{k1} & \cdots & \sum_k A_{mk}B_{kn} \end{bmatrix} \\
&= AB^T + AB \qquad (1.65)
\end{aligned}$$

If B is symmetric, as it often is in partial derivatives of the form above, then this can be simplified to

$$\frac{\partial \mathrm{Tr}(ABA^T)}{\partial A} = 2AB \qquad \text{if } B = B^T \qquad (1.66)$$

A number of additional interesting results related to matrix calculus can be found in [Ske98, Appendix B].

1.1.4 The history of matrices

This section is a brief diversion to present some of the history of matrix theory. Much of the information in this section is taken from [OCo96].

The use of matrices can be found as far back as the fourth century BC. We see in ancient clay tablets that the Babylonians studied problems that led to simultaneous linear equations. For example, a tablet dating from about 300 BC contains the following problem: "There are two fields whose total area is 1800 units. One produces grain at the rate of 2/3 of a bushel per unit while the other produces grain at the rate of 1/2 a bushel per unit. If the total yield is 1100 bushels, what is the size of each field?"

Later, the Chinese came even closer to the use of matrices. In [She99] (originally published between 200 BC and 100 AD) we see the following problem: "There are three types of corn, of which three bundles of the first, two of the second, and one of the third make 39 measures. Two of the first, three of the second, and one of the third make 34 measures. And one of the first, two of the second and three of the third make 26 measures. How many measures of corn are contained in one bundle of each type?" At that point, the ancient Chinese essentially use Gaussian elimination (which was not well known until the 19th century) to solve the problem.

In spite of this very early beginning, it was not until the end of the 17th century that serious investigation of matrix algebra began. In 1683, the Japanese