Univ. of Washington

Lecture : Vector & Matrix Derivatives

Lecturer: Dan Calderone

Vector Derivatives

Derivatives are linear maps that convert perturbations in function arguments into perturbations in the function themselves. Consider $x \in \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$. $f(x)$ is a scalar. The derivative $\frac{\partial f}{\partial x}$ is the row vector

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

such that

$$\Delta f \approx \frac{\partial f}{\partial x}\Delta x = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} \begin{bmatrix} \Delta x_1 \\ \vdots \\ \Delta x_n \end{bmatrix} \tag{1}$$

where $\Delta f \in \mathbb{R}$ and $\Delta x \in \mathbb{R}^n$ are perturbations in f and x, respectively. Note that if f is linear, ie. $f(x) = b^\top x$, then $\frac{\partial f}{\partial x} = b^\top$. Note that the perturbation form in (1) can be useful in computing vector derivatives in tricky situations. For example, suppose $f(x) = x^\top Q x + b^\top x$. In order to compute the derivative, we can perturb each instance of x separately and add up the perturbations. (The ability to perturb each instance of x separately is called the *product rule*.) Then we rearrange the right hand side (RHS) into the form of (1).

$$\Delta f = \Delta x^\top Q x + x^\top Q \Delta x + b^\top \Delta x \tag{2}$$

Noticing that each of the terms in the RHS is a scalar, we can transpose as necessary.

$$\begin{aligned} \Delta f &= (\Delta x^\top Q x)^\top + x^\top Q \Delta x + b^\top \Delta x && (3)\\ &= \left(x^\top (Q + Q^\top) + b^\top\right)\Delta x && (4)\\ \Rightarrow \qquad \frac{\partial f}{\partial x} &= x^\top (Q + Q^\top) + b^\top && (5) \end{aligned}$$

Now suppose $f(x)$ is a vector valued function $f : \mathbb{R}^n \to \mathbb{R}^n$. The derivative is now an $m \times n$ matrix

$$\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \tag{6}$$

such that

$$\Delta f = \begin{bmatrix} \Delta f_1 \\ \vdots \\ \Delta f_m \end{bmatrix} \approx \frac{\partial f}{\partial x} \Delta x = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m}{\partial x_1} & \cdots & \frac{\partial f_m}{\partial x_n} \end{bmatrix} \begin{bmatrix} \Delta x_1 \\ \vdots \\ \Delta x_n \end{bmatrix} \tag{7}$$

where $\Delta f \in \mathbb{R}^m$ and $\Delta x \in \mathbb{R}^n$. Note that when $\frac{\partial f}{\partial x}$ is a matrix it is referred to as a *Jacobian*.

Now suppose we have a scalar function $f(x)$ and we want to compute its second derivative. Differentiating once gives

$$\frac{\partial f}{\partial x} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \cdots & \frac{\partial f}{\partial x_n} \end{bmatrix} \tag{8}$$

Now treating $\frac{\partial f}{\partial x}$ as a vector valued function, we can compute the second derivative

$$\frac{\partial^2 f}{\partial x^2} = \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \tag{9}$$

The matrix $\frac{\partial^2 f}{\partial x^2}$ is symmetric since $\frac{\partial^2 f}{\partial x_i \partial x_j} = \frac{\partial^2 f}{\partial x_j \partial x_i}$ and is referred to as the *Hessian* of the function $f(x)$. Second derivatives are used to approximate perturbations of first derivatives

$$\Delta \frac{\partial f}{\partial x} \approx \Delta x^\top \frac{\partial^2 f}{\partial x^2} = \begin{bmatrix} \Delta x_1 \\ \vdots \\ \Delta x_n \end{bmatrix}^\top \begin{bmatrix} \frac{\partial^2 f}{\partial x_1^2} & \cdots & \frac{\partial^2 f}{\partial x_1 \partial x_n} \\ \vdots & & \vdots \\ \frac{\partial^2 f}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2 f}{\partial x_n^2} \end{bmatrix} \tag{10}$$

For the quadratic function $f(x) = x^\top Q x + b^\top x$, we can use the perturbative perspective to compute

$$\Delta \frac{\partial f}{\partial x} = \Delta x^\top \frac{\partial^2 f}{\partial x^2} = \Delta x^\top (Q + Q^\top) \Rightarrow \qquad \frac{\partial^2 f}{\partial x^2} = Q + Q^\top \tag{11}$$

Note that often we write $\frac{\partial^2 f}{\partial x^2} = 2Q$. This is consistent with above formula assuming that $Q = Q^\top$ is symmetric. Any time we consider a quadratic form $x^\top Q x$, we assume that Q is symmetric. The reason for this is that if it's not symmetric, only the symmetric part of it affects the product $x^\top Q x$. Explicitly, write

$$\begin{aligned} x^\top Q x &= x^\top \Big(\tfrac{1}{2}\big(Q + Q^\top\big) + \tfrac{1}{2}\big(Q - Q^\top\big) \Big) x \\ &= \tfrac{1}{2} x^\top \big(Q + Q^\top\big) x + \tfrac{1}{2} x^\top \big(Q - Q^\top\big) x \\ &= \tfrac{1}{2} x^\top \big(Q + Q^\top\big) x + \underbrace{\tfrac{1}{2} x^\top Q x - \tfrac{1}{2} x^\top Q^\top x}_{=0} \end{aligned}$$

The first part of the expansion is the symmetric part of Q. The second part is the skew symmetric part and $x^\top K x = 0$ for any $K = -K^\top$ (K is skew-symmetric).

Using this structure, we also comment on how to express a vector valued Taylor expansion. Up to the quadratic term a Taylor expansion for $f(x)$ around a point x_0 is given by

$$f(x) = f(x_0) + \left.\frac{\partial f}{\partial x}\right|_{x_0} \Delta x + \Delta x^\top \left.\frac{\partial^2 f}{\partial x^2}\right|_{x_0} \Delta x + \cdots \quad \text{where} \quad \Delta x = x - x_0$$

Note how this relates to the perturbation analysis ideas discussed above.

Chain Rule

One important practical tool for taking vector derivatives is the *chain rule*. One of the reasons to be careful about how we arrange vector derivatives, and particularly to write the derivative of a function $f : \mathbb{R}^n \to \mathbb{R}^m$ as an $m \times n$ matrix $\frac{\partial f}{\partial x} \in \mathbb{R}^{m \times n}$ is so that it is easy to apply the chain rule consistent with the rules of multiplication. Specifically, consider several functions

$$h(z) : \mathbb{R}^q \to \mathbb{R}^m, \qquad g(y) : \mathbb{R}^p \to \mathbb{R}^q, \qquad f(x) : \mathbb{R}^n \to \mathbb{R}^p$$

The derivatives of each function are matrices

$$\frac{\partial h}{\partial z} \in \mathbb{R}^{m \times q}, \qquad \frac{\partial g}{\partial y} \in \mathbb{R}^{q \times p}, \qquad \frac{\partial f}{\partial x} \in \mathbb{R}^{p \times n}$$

Suppose these functions are now composed together $u(x) = h\big(g\big(f(x)\big)\big)$. The derivative of $u(x)$ with respect to x can then be computed as

$$\frac{\partial u}{\partial x} = \frac{\partial}{\partial x}\Big(h\big(g\big(f(x)\big)\big)\Big) = \left[\frac{\partial h}{\partial z}\right]\left[\frac{\partial g}{\partial y}\right]\left[\frac{\partial f}{\partial x}\right]$$

Carefully note the order of the vector derivative matrices and also how the dimensions of each matrix match up for the matrix multiplication to work. Note also how our perturbation analysis goes through.

$$\Delta u = \frac{\partial u}{\partial x}\Delta x = \left[\frac{\partial h}{\partial z}\right] \underbrace{\left[\frac{\partial g}{\partial y}\right] \underbrace{\left[\frac{\partial f}{\partial x}\right]\Delta x}_{\Delta y}}_{\Delta z}$$

To be completely accurate we have to be careful to plug in the correct argument to each derivative matrix and thus we should write

$$\frac{\partial u}{\partial x} = \left.\frac{\partial h}{\partial z}\right|_{g(h(x))} \left.\frac{\partial g}{\partial y}\right|_{f(x)} \left.\frac{\partial f}{\partial x}\right|_{x}$$

As an example consider the function $u(x) = e^{-\frac{1}{2}y^\top Qy}$ where $y = Hx$ for $y \in \mathbb{R}^p$ and $H \in \mathbb{R}^{p\times n}$. (This is essentially the equation for a slice of a multivariate Gaussian.). Here we can take

$$h(z) = e^z, \qquad g(y) = -\tfrac{1}{2}y^\top Qy, \qquad f(x) = Hx$$

with derivatives

$$\frac{\partial h}{\partial z} = e^z \in \mathbb{R}^{1\times 1}, \qquad \frac{\partial g}{\partial y} = -\tfrac{1}{2}y^\top(Q+Q^\top) \in \mathbb{R}^{1\times p}, \qquad \frac{\partial f}{\partial x} = H \in \mathbb{R}^{p\times n}$$

Plugging in $y = Hx$ and $z = -\frac{1}{2}y^\top Qy$ gives

$$\begin{aligned}\frac{\partial u}{\partial x} &= -e^z\Big|_{-\frac{1}{2}x^\top H^\top QHx}\ \tfrac{1}{2}y^\top(Q+Q^\top)\Big|_{Hx} H \\ &= -\tfrac{1}{2}e^{-\frac{1}{2}x^\top H^\top QHx}x^\top H^\top \tfrac{1}{2}\big(Q+Q^\top\big)H \\ &= -\tfrac{1}{2}e^{-\frac{1}{2}x^\top H^\top QHx}x^\top H^\top QH\end{aligned}$$

where in the last line we've assumed Q is symmetric. Note carefully how all the dimensions work out so that the above expression is consistent with the rules of matrix multiplication. Again, the fact that the dimensions work out is not a fluke but rather because we were careful to be consistent with our definition of derivatives and application of the chain rule.

Matrix Derivatives

We now consider taking derivatives of functions $F(X)$ where either the input X or the output F are matrices. The perturbation analysis from above works exactly the same, but these are generally trickier to write down because they are usually higher (more than two) dimensional tensors. The one exception which we will deal with first is when either X or F is simply a scalar.

We start with the case where X is a scalar, $F : \mathbb{R} \to \mathbb{R}^{m\times n}$. In this case, we will usually define

$$\frac{\partial F}{\partial X} = \begin{bmatrix} \frac{\partial F_{11}}{\partial X} & \cdots & \frac{\partial F_{1n}}{\partial X} \\ \vdots & & \vdots \\ \frac{\partial F_{m1}}{\partial X} & \cdots & \frac{\partial F_{mn}}{\partial X} \end{bmatrix} \tag{12}$$

The perturbation analysis can then be written as

$$\Delta F \approx= \frac{\partial F}{\partial X}\Delta X = \begin{bmatrix} \frac{\partial F_{11}}{\partial X} & \cdots & \frac{\partial F_{1n}}{\partial X} \\ \vdots & & \vdots \\ \frac{\partial F_{m1}}{\partial X} & \cdots & \frac{\partial F_{mn}}{\partial X} \end{bmatrix} \Delta X$$

Each element of $\frac{\partial F}{\partial X}$ is simply the scalar derivative of the corresponding element of F with respect to X. $\left[\frac{\partial F}{\partial X}\right]_{ij} = \frac{\partial F_{ij}}{\partial X}$.

We now consider the case where F is a scalar and X is a matrix, $F : \mathbb{R}^{m\times n} \to \mathbb{R}$. In this case,

$$\frac{\partial F}{\partial X} = \begin{bmatrix} \frac{\partial F}{\partial X_{11}} & \cdots & \frac{\partial F}{\partial X_{ij}} \\ \vdots & & \vdots \\ \frac{\partial F}{\partial X_{m1}} & \cdots & \frac{\partial F}{\partial X_{mn}} \end{bmatrix} \tag{13}$$

Note the similarities and differences with (12). The perturbation analysis will involve summing over all elements of $\frac{\partial F}{\partial X}$. We could, for example, write

$$\Delta F = \sum_{ij} \frac{\partial F}{\partial X_{ij}} \Delta X_{ij}$$

However, in many practical problems, scalar functions of matrices are written in terms of quadratic forms or trace operators such as $F(X) = a^\top X b$ with $a \in \mathbb{R}^m$, $b \in \mathbb{R}^n$ or $F(X) = \text{Tr}(C^\top X)$ with $C \in \mathbb{R}^{m\times n}$. It is worth knowing how to deal with these cases specially. Our analysis will leverage properties of the trace operator and Euclidean matrix inner product $\langle C, X\rangle = \text{Tr}(C^\top X)$.

Paralleling the notation of a vector dot product, the basic inner product on the space of matrices is

$$\langle Y, X\rangle = \sum_{ij} Y_{ij} X_{ij} = \text{Tr}(Y^\top X)$$

Here we simply match up the corresponding elements of Y and X and sum over them. One can check that the final expression $\text{Tr}(Y^\top X)$ does exactly this. (In practice, one would not compute the full product $Y^\top X$ in order to calculate this inner product cause only the diagonal is needed, but it is quite useful for analytic purposes.). Using this inner product idea, we can rewrite our perturbation analysis as

$$\Delta F = \sum_{ij} \frac{\partial F}{\partial X_{ij}} \Delta X_{ij} = \left\langle \frac{\partial F}{\partial X}, \Delta X \right\rangle = \text{Tr}\left(\frac{\partial F}{\partial X}^\top \Delta X \right) \tag{14}$$

Again, this can be a useful way to think of $\frac{\partial F}{\partial X}$, it is the matrix object that if we take the matrix inner product of it with a perturbation ΔX then we get the perturbation in F, ΔF. The trace expression can also be quite useful because it is often easy to write our function $F(X)$ in a form that looks like the far RHS of (14). We give several examples. The function $F(X) = \text{Tr}(C^\top X)$ is in this form already and we immediately have that

$$F(X) = \text{Tr}(C^\top X) \quad \implies \quad \frac{\partial F}{\partial X} = C$$

The function $F(X) = a^\top X b$ is a little trickier, but since it is a scalar value we can put it inside a trace operator without changing it, ie. $F(X) = \text{Tr}(F(X)) = \text{Tr}(a^\top X b)$. (This is always possible for any scalar function F). We can then leverage the cyclic property of traces.

$$\text{Tr}(ABCD) = \text{Tr}(DABC) = \text{Tr}(CDAB) = \text{Tr}(BCDA)$$

assuming that ABC was square in the first place. (It is worth playing around with this formula and convincing yourself that is true as well as seeing how the dimensions of A, B, C, D come into play. The only requirement for this to work is that $ABCD$ is square (and that the dimensions of A, B, C, D are compatible for the original multiplication.) For this reason, trace algebra is actually quite pleasant because you can change the order of matrices in a product (which is not possible when the product is not inside a trace). Returning to our original formula we can write

$$F(X) = a^\top Xb = \text{Tr}(a^\top Xb) = \text{Tr}(ba^\top X) \quad \Longrightarrow \quad \frac{\partial F}{\partial X} = ab^\top$$

Similarly for $F(X) = \text{Tr}\big(AXB\big)$, we can write

$$F(X) = \text{Tr}(AXB) = \text{Tr}(BAX) \quad \Longrightarrow \quad \frac{\partial F}{\partial X} = A^\top B^\top$$

In any practical setting where one is taking derivatives with respect to matrices, being able to use these algebraic tricks involving traces is crucial. Trying to compute out each element of (13) individually and then organize them back into a usable expression is not doable.

We now give a more complicated example both for practice and also to illustrate a practical method for computing derivatives that is widely applicable and quite powerful in many cases. Often the best way to compute a derivative is to do a perturbation analysis, cancel zero-order and higher-order terms, and then read off the derivative from this expansion. Practically speaking for difficult matrix or vector derivatives this can be a quite powerful technique. To illustrate this procedure, we will do it on the function

$$f(X) = a^\top X^\top (A + XB)^{-1} Xb$$

where $X \in \mathbb{R}^{m \times n}$, $A \in \mathbb{R}^{m \times m}$, $B \in \mathbb{R}^{n \times m}$, and $a \in \mathbb{R}^m$ and $b \in \mathbb{R}^n$. We start by plugging in $X + \Delta X$ in for X in the above expression.

$$f(X + \Delta X) = a^\top (X + \Delta X)^\top (A + (X + \Delta X)B)^{-1} (X + \Delta X)b$$

Using the Woodbury matrix identity, we get

$$\begin{aligned} f(X + \Delta X) = a^\top &\Big(X + \Delta X\Big)^\top \\ &\cdot \Big(\big(A + XB\big)^{-1} - \big(A + XB\big)^{-1} \Delta X (I + B\big(A + XB\big)^{-1} \Delta X)^{-1} B \big(A + XB\big)^{-1}\Big) \\ &\cdot \Big(X + \Delta X\Big)b \end{aligned}$$

and then expanding again gives

$$
\begin{aligned}
f(X+\Delta X) = a^\top X^\top \big(A+XB\big)^{-1} Xb \\
&+ a^\top \Delta X^\top \big(A+XB\big)^{-1} Xb \\
&- a^\top X^\top \big(A+XB\big)^{-1} \Delta X (I + B\big(A+XB\big)^{-1}\Delta X)^{-1} B\big(A+XB\big)^{-1} Xb \\
&- a^\top \Delta X^\top \big(A+XB\big)^{-1} \Delta X (I + B\big(A+XB\big)^{-1}\Delta X)^{-1} B\big(A+XB\big)^{-1} Xb \\
&+ a^\top X^\top \big(A+XB\big)^{-1} \Delta X b \\
&+ a^\top \Delta X^\top \big(A+XB\big)^{-1} \Delta X b \\
&- a^\top X^\top \big(A+XB\big)^{-1} \Delta X (I + B\big(A+XB\big)^{-1}\Delta X)^{-1} B\big(A+XB\big)^{-1} \Delta X b \\
&- a^\top \Delta X^\top \big(A+XB\big)^{-1} \Delta X (I + B\big(A+XB\big)^{-1}\Delta X)^{-1} B\big(A+XB\big)^{-1} \Delta X b
\end{aligned}
$$

We now want to isolate the first order perturbation terms. First, we use the fact that $f(X) = a^\top X^\top (A+XB)^{-1} Xb$ and that $(I + B(A+XB)^{-1}\Delta X) \to I$ as $||\Delta X||_2 \to 0$. Plugging these in and reorganizing gives

$$
\begin{aligned}
f + \Delta f \approx f(X) & && \text{(zero-order)} \\
&+ a^\top \Delta X^\top \big(A+XB\big)^{-1} Xb && \text{(first-order)} \\
&+ a^\top X^\top \big(A+XB\big)^{-1} \Delta X b \\
&- a^\top X^\top \big(A+XB\big)^{-1} \Delta X B \big(A+XB\big)^{-1} Xb \\
&- a^\top \Delta X^\top \big(A+XB\big)^{-1} \Delta X B \big(A+XB\big)^{-1} Xb && \text{(second-order)} \\
&+ a^\top \Delta X^\top \big(A+XB\big)^{-1} \Delta X b \\
&- a^\top X^\top \big(A+XB\big)^{-1} \Delta X B \big(A+XB\big)^{-1} \Delta X b \\
&- a^\top \Delta X^\top \big(A+XB\big)^{-1} \Delta X B \big(A+XB\big)^{-1} \Delta X b && \text{(third-order)}
\end{aligned}
$$

Canceling the zero-order, second-order, and third-order terms leaves the first order terms.

$$
\begin{aligned}
\Delta f \approx a^\top \Delta X^\top \big(A+XB\big)^{-1} Xb \\
&+ a^\top X^\top \big(A+XB\big)^{-1} \Delta X b \\
&- a^\top X^\top \big(A+XB\big)^{-1} \Delta X B \big(A+XB\big)^{-1} Xb
\end{aligned}
$$

We now want to rearrange this equation into a form where we can read off the derivative. Since

these terms are scalars, we can transposes, apply traces, and rearrange terms to get

$$\begin{aligned}\Delta f &\approx \text{Tr}\Big(a^\top \Delta X^\top \big(A+XB\big)^{-1} Xb \\ &\quad + a^\top X^\top \big(A+XB\big)^{-1} \Delta X b - a^\top X^\top \big(A+XB\big)^{-1} \Delta X B \big(A+XB\big)^{-1} Xb\Big) \\ &\approx \text{Tr}\Bigg(\Big(ab^\top X^\top \big(A+XB\big)^{-\top} + ba^\top X^\top \big(A+XB\big)^{-1} \\ &\quad - B\big(A+XB\big)^{-1} Xba^\top X^\top \big(A+XB\big)^{-1}\Big)\Delta X\Bigg)\end{aligned}$$

Using the equation $\Delta f = \text{Tr}\left(\frac{\partial f}{\partial X}^\top \Delta X\right)$, we have that

$$\frac{\partial f}{\partial X}^\top = ab^\top X^\top \big(A+XB\big)^{-\top} + ba^\top X^\top \big(A+XB\big)^{-1} - B\big(A+XB\big)^{-1} Xba^\top X^\top \big(A+XB\big)^{-1}$$