Mathematics for Economists
Aalto BIZ
Spring2022
Juuso Välimäki

Derivative as linear approximation

Linear functions

A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is said to be *linear* if the following two conditions are satisfied:

i) (Homogeneity) For all $\lambda \in \mathbb{R}$ and for all $\boldsymbol{x} \in \mathbb{R}^n$, $f(\lambda \boldsymbol{x}) = \lambda f(\boldsymbol{x})$,

ii) (Additivity) For all $\boldsymbol{x}, \boldsymbol{y} \in \mathbb{R}^n$, $f(\boldsymbol{x} + \boldsymbol{y}) = f(\boldsymbol{x}) + f(\boldsymbol{y})$.

By taking $\lambda = 0$ in i), we see that $f(0) = 0$ for all linear functions. In the case of $n = 1$, this rules out functions whose graphs are straight lines that do not go through the origin. In this simplest setting, i) actually implies ii), and fixing $f(1)$ determines the entire function.

For $n > 1$, requirement ii) has bite. Observe that we can write $\boldsymbol{x} = \sum_{i=1}^n x_i \boldsymbol{e}^i$. By i), $f(x_i \boldsymbol{e}^i) = x_i f(\boldsymbol{e}^i)$ for all i, x_i. By ii),

$$f(\boldsymbol{x}) = f(\sum_{i=1}^n x_i \boldsymbol{e}^i) = \sum_{i=1}^n x_i f(\boldsymbol{e}^i).$$

Hence a linear function is completely determined by n values $f(\boldsymbol{e}^i)$. If $m = 1$, then $f(\boldsymbol{e}^i) \in \mathbb{R}$ for all i and letting $f(\boldsymbol{e}^i) = a_i$ we see that all real linear functions from $\mathbb{R}^n$ are given by inner products $\boldsymbol{a} \cdot \boldsymbol{x} = \sum_{i=1}^n a_i x_i$.

If $m > 1$, then each $f(\boldsymbol{e}^i)$ is an m -dimensional vector. If we denote $\boldsymbol{a}^i = f(\boldsymbol{e}^i) \in \mathbb{R}^m$, we have as before $f(x) = \sum_{i=1}^n x_i \boldsymbol{a}^i$. Writing $\boldsymbol{A} = [\boldsymbol{a}^1, ..., \boldsymbol{a}^n]$ for the matrix consisting of columns $\boldsymbol{a}^i$. But this means that

$$f(x) = \boldsymbol{A}\boldsymbol{x}.$$

Many of the properties of linear functions also extend to *affine* functions of the form

$$f(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x} + \boldsymbol{b},$$

for some $\boldsymbol{b} \in \mathbb{R}^m$. Actually this is not so bad because by shifting the origin to $(0, f(0))$, $\hat{f}((\boldsymbol{x}) := f(\boldsymbol{x}) - f(0) = \boldsymbol{A}\boldsymbol{x}$ is linear.

Why are linear functions so much simpler than non-linear? i) A change in $\boldsymbol{x}$ has the same effect regardless of the starting point:

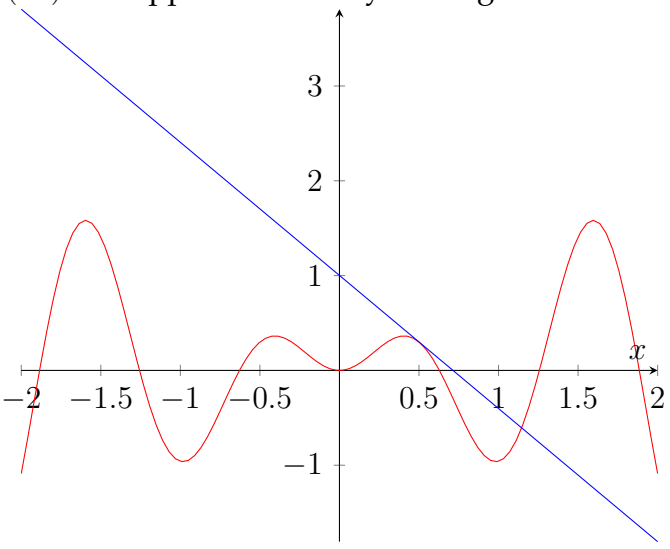
$$f(\boldsymbol{x}) - f(\hat{\boldsymbol{x}}) = A(\boldsymbol{x} - \hat{\boldsymbol{x}}).$$

ii) A function $f : \mathbb{R}^n \to \mathbb{R}^m$ is said to be *surjective* (or *onto*) if for all $\boldsymbol{b} \in \mathbb{R}^m$, there is an $\boldsymbol{x} \in \mathbb{R}^n$ such that $f(\boldsymbol{x}) = \boldsymbol{b}$. f is said to be *injective* or *one-to-one* if for all $\boldsymbol{x} \neq \boldsymbol{x}'$, $\quad f(\boldsymbol{x}) \neq f(\boldsymbol{x}')$. f is said to be *bijective* if it is injective and surjective. Bijective functions f have an *inverse function* $f^{-1} : \mathbb{R}^m \to /R^n$ such that $f^{-1}(f(\boldsymbol{x})) = \boldsymbol{x}$ and $f(f^{-1}(\boldsymbol{y})) = \boldsymbol{y}$. In matrix algebra, we saw that if $f(\boldsymbol{x}) = \boldsymbol{A}\boldsymbol{x}$, then f is bijective if and only if $\boldsymbol{A}$ has full rank. Gaussian elimination (or the determinant) gives an easy way of determining when linear functions are bijective and computing the inverse function $f^{-1}(\boldsymbol{x}) = A^{-1}\boldsymbol{x}$.

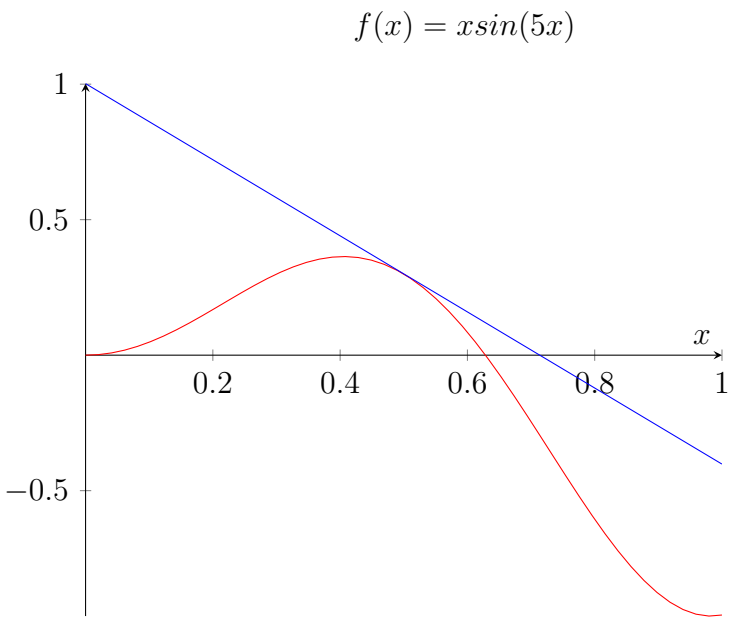
The derivative: Inducing a linear approximation

Motivation

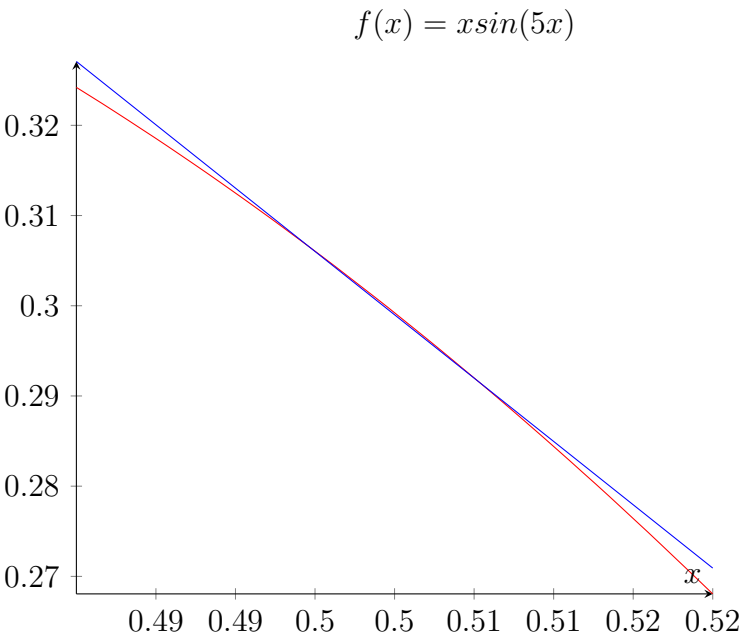
Consider the graph of the following highly non-linear function $f(x) = xsin(5x)$ and approximate it by its tangent at $x = 0.5$. Not a great success:



It is a bit less variable over the interval $[0, 1]$:



On the interval $[0.48, 0.52]$, it looks almost linear:



As we zoom closer to a fixed point (here $\hat{x} = 0.5$), the original highly irregular function starts resembling a linear function. The point of computing derivatives is exactly this: we want to get good linear approximations to non-linear functions near a given point $(\hat{x}, f(\hat{x}))$ on the graph of f.

Real valued functions of a single variable

From elementary calculus, the derivative of $f : \mathbb{R} \to \mathbb{R}$ at $\hat{x} \in \mathbb{R}$ is defined as:

$$Df(\hat{x}) = f'(\hat{x}) = \frac{\mathrm{d}f(x)}{\mathrm{d}x}\bigg|_{x=\hat{x}} = \lim_{h \to 0} \frac{f(\hat{x}+h) - f(\hat{x})}{h}.$$

You may recall that this numerical value (whenever the limit exists) can be interpreted as the slope of the tangent to f at $\hat{x}$.

This is of course fine, but an alternative way to think about the derivative at $\hat{x}$ is as a linear approximation through origin $(\hat{x}, f(\hat{x}))$ that best approximates f near $\hat{x}$. If the limit exists, we can write:

$$f(\hat{x}+h) - f(\hat{x}) = Df(\hat{x})h + \varepsilon(h),$$

where $\lim_{h \to 0} \frac{\varepsilon(h)}{h} = 0$. Sometimes, the reminder $\varepsilon(h)$ is just written as higher order terms or h.o.t.

The reason for insisting on this seemingly trivial point is that this view of the derivative generalizes immediately to multivariate functions $f : \mathbb{R}^n \to \mathbb{R}^m$. It also tells you immediately that for small h you get a good approximations for the change in the value $f(\hat{x}+h) - f(\hat{x}) = Df(\hat{x})h$. Notice also that for functions of a single variable, there is only a single direction h that we need to consider.

At this point, it may be a good idea to refresh the rules for computing derivatives (a handout on this is in the materials for week 2).

Real valued functions of a multiple variables

Consider next real valued functions defined in the $x - y$ plane, in other words functions $f : \mathbb{R}^2 \to \mathbb{R}$. Since we need a third dimension to graph the value corresponding to the value of the function at each point of the plane, we run into some difficulties in representing a function on a two-dimensional screen.

We have some options available:

1. Drawing a 3-d graph:

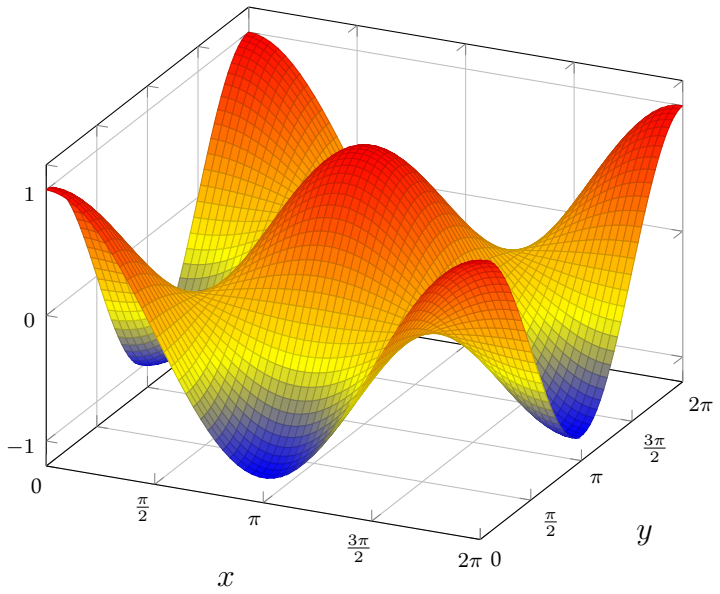


Figure 1: The graph of $f(x, y) = \cos(x)\cos(y)$.

2. Drawing 2-d slices of the 3-d graph:

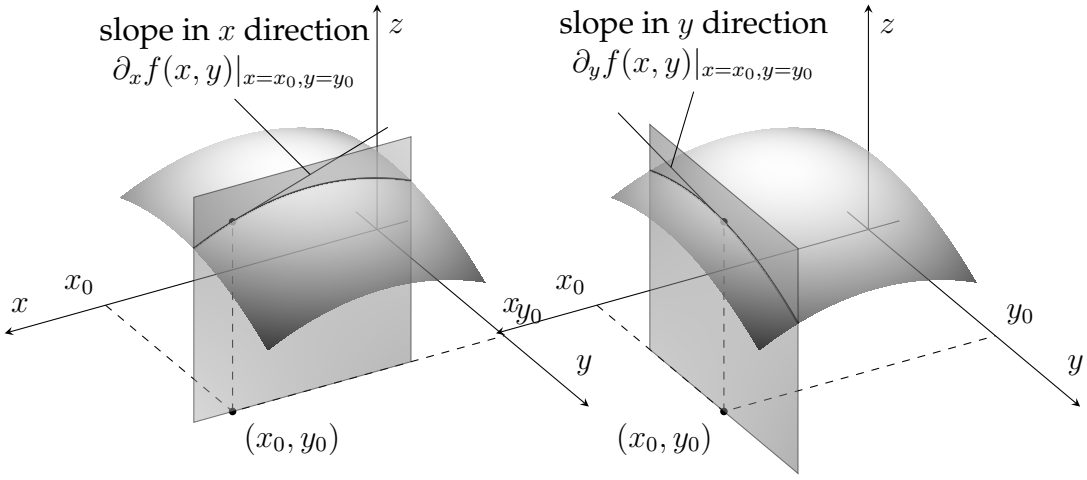


Figure 2: Cross sections of f in the x and y direction at (x_0, y_0).

3. Drawing level curves of the function in its domain:

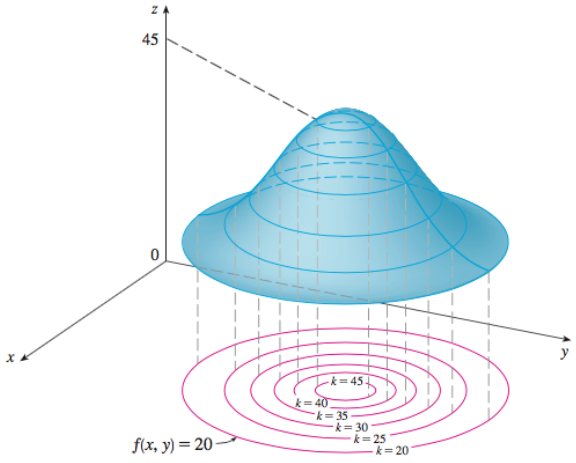


Figure 3: Some level curves of f.

4. combining all three ways in one picture:

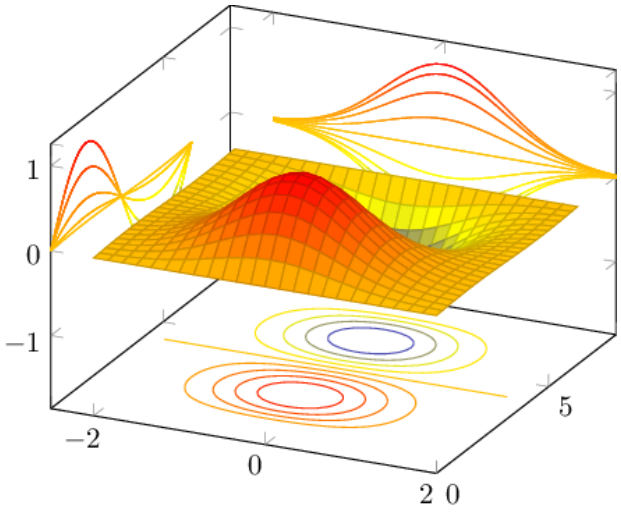


Figure 4: The graph of f together with some of its cross sections and level curves.

Overall goal: Linear approximation

For functions $f : \mathbb{R}^n \to \mathbb{R}$ variables, a good approximation does well in all n directions in the domain of f. Ideally we would end up with a formula that allows us to approximate f at $\hat{\boldsymbol{x}} \in \mathbb{R}^n$ with a linear function. Recall that a linear function $f : \mathbb{R}^n \to \mathbb{R}$ take the form of an inner product (multiplication by a row vector or a $(1 \times n)$ matrix. So we would like to have for some $\boldsymbol{a} \in \mathbb{R}^n$:

$$f(\hat{\boldsymbol{x}} + \Delta\boldsymbol{x}) - f(\hat{\boldsymbol{x}}) = \boldsymbol{a} \cdot \Delta\boldsymbol{x} + \text{ h.o.t. },$$

whenever $\Delta\boldsymbol{x}$ is small.

If we can find such a $\boldsymbol{a}$, then computing the approximation in all possible directions is simply a matter of taking inner products of vectors and we are back to linear algebra as long as we stay close to the original point $\hat{\boldsymbol{x}}$.

How to get there: Partial Derivatives

Recall that we denote unit vectors in the standard coordinate system of $\mathbb{R}^n$ by $\boldsymbol{e}^i$. Moving away from $\hat{\boldsymbol{x}}$ in the direction of the first coordinate axis is then denoted by moving to $\hat{\boldsymbol{x}} + \Delta\boldsymbol{x}$ with $\Delta\boldsymbol{x} = h\boldsymbol{e}^1$ for some $h \in \mathbb{R}$. But this means that all the other components in $\boldsymbol{x}$ remain fixed.

When we consider changes in the direction of a coordinate axis, we are really analyzing a function of a single real variable since all the other components in $\hat{\boldsymbol{x}}$ do not change. But this means that we can compute an approximation for f in the direction of $\boldsymbol{e}^i$ by computing the derivative:

$$\lim_{h \to 0} \frac{f(\hat{\boldsymbol{x}} + h\boldsymbol{e}^i) - f(\hat{\boldsymbol{x}})}{h},$$

exactly as we did before since we now have a function of a single variable x_i the x_j for $j \neq i$ are 'fixables'.

We call this limit the partial derivative of f at $\hat{\boldsymbol{x}}$ and denote it by:

$$D_{x_i} f(\hat{\boldsymbol{x}}) :=: \frac{\partial f(\hat{\boldsymbol{x}})}{\partial x_i} := \lim_{h \to 0} \frac{f(\hat{\boldsymbol{x}} + h\boldsymbol{e}^i) - f(\hat{\boldsymbol{x}})}{h}.$$

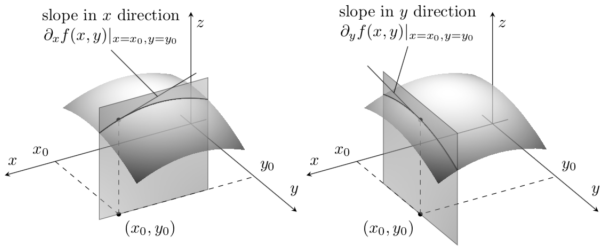


Figure 5: Partial derivatives of f at (x_0, y_0).

But this is all we need if a linear approximation exists! A linear approximation in the direction $\Delta \boldsymbol{x} = \boldsymbol{e}^i$ must coincide with $\frac{\partial f(\hat{\boldsymbol{x}})}{\partial x_i}$. But each direction $\Delta \boldsymbol{x}$ can be written as:

$$\Delta \boldsymbol{x} = \sum_{i=1}^{n} x_i \boldsymbol{e}^i,$$

so by linearity we get for all $\Delta \boldsymbol{x} = h(x_1, ..., x_n)$ that

$$f(\hat{\boldsymbol{x}} + \Delta \boldsymbol{x}) - f(\hat{\boldsymbol{x}}) = D_{\boldsymbol{x}} f(\hat{\boldsymbol{x}}) \cdot \Delta \boldsymbol{x} + \text{ h.o.t. },$$

where $D_{\boldsymbol{x}} f(\hat{\boldsymbol{x}})$ is the row vector of partial derivatives

$$D_{\boldsymbol{x}} f(\hat{\boldsymbol{x}}) = (\frac{\partial f(\hat{\boldsymbol{x}})}{\partial x_1}, ..., \frac{\partial f(\hat{\boldsymbol{x}})}{\partial x_n}).$$

Figure 6 shows a plane approximating a two-dimensional surface in three-dimensional space

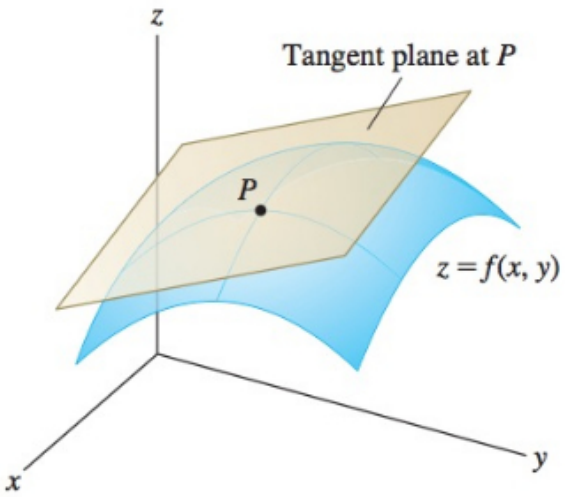


Figure 6: Linear approximation to f at point P.

Extra material: Existence of linear approximation

It can be shown that such a linear approximation exists if all the partial derivatives evaluated at $\boldsymbol{x}$ are continuous functions of $\boldsymbol{x}$, i.e. the point at which they are evaluated.

The mere existence of partial derivatives is not enough to guarantee even continuity of the function. For an example, you can consider the function around $(x, y) = (0, 0)$:

$$f(x,y) = \begin{cases} \frac{xy}{x^2+y^2} & \text{if} \quad (x,y) \neq (0,0), \\ 0 & \text{if} \quad (x,y) = (0,0). \end{cases}$$

The requirement of continuous partial derivatives obviously fails in this case. We will not prove the intuitive result that with continuous partial derivatives, you get the existence of the linear approximation. In this case, we say that f is differentiable at $\hat{\boldsymbol{x}}$. The proof can be found in any book on advanced calculus.

Vector valued functions of multiple real variables

A function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ is a vector of functions $f_i : \mathbb{R}^n \to \mathbb{R}$:

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{pmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_m(\boldsymbol{x}) \end{pmatrix}.$$

Since each of the component functions is a real valued function of n variables, we have from the previous subsection that a linear approximation of f_i at $\hat{\boldsymbol{x}}$ is given by the derivative $D_{\boldsymbol{x}} f(\hat{\boldsymbol{x}})$. If all the partial derivatives of all component functions exist and are continuous at $\hat{\boldsymbol{x}}$, then the derivative of $\boldsymbol{f}$ at $\hat{\boldsymbol{x}}$ is the $m \times n$ matrix:

$$D_{\boldsymbol{x}} \boldsymbol{f}(\boldsymbol{x}) = \begin{pmatrix} \frac{\partial f_1(\hat{\boldsymbol{x}})}{\partial x_1} & \cdots & \frac{\partial f_1(\hat{\boldsymbol{x}})}{\partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial f_m(\hat{\boldsymbol{x}})}{\partial x_1} & \cdots & \frac{\partial f_m(\hat{\boldsymbol{x}})}{\partial x_n} \end{pmatrix}.$$

Since partial derivatives can be viewed as standard derivatives in a fixed direction, the rules for computing derivatives remain valid for multivariate functions. In particular, we have the chain rule for $h(\boldsymbol{x}) := f(g(\boldsymbol{x}))$, where $f : \mathbb{R}^k \to \mathbb{R}^m$ and $g : \mathbb{R}^n \to \mathbb{R}^k$ Let $\hat{\boldsymbol{y}} = (\hat{y}_1, ..., \hat{y}_k) = (g_1(\hat{\boldsymbol{x}}), ..., g_1(\hat{\boldsymbol{x}}))$:

$$D_{\boldsymbol{x}} h(\boldsymbol{x})) = D_{\boldsymbol{y}} f(\hat{\boldsymbol{y}}) D_{\boldsymbol{x}} g(\hat{\boldsymbol{x}}).$$

Writing this matrix multiplication explicitly gives the ij^{th} element of $D_{\boldsymbol{x}} h(\boldsymbol{x})$ as:

$$\frac{\partial h_i(\hat{\boldsymbol{x}})}{\partial x_j} = \sum_{i=1}^{k} \frac{\partial f_i(\hat{\boldsymbol{y}})}{\partial y_k} \frac{\partial g_k(\hat{\boldsymbol{x}})}{\partial x_j}.$$

To recap: the derivative of $\boldsymbol{f}$ at $\hat{\boldsymbol{x}}$ is the matrix of its partial derivatives evaluated at that point. This matrix (of fixed numbers) generates the linear function

$$\boldsymbol{f}(\hat{\boldsymbol{x}} + \Delta\boldsymbol{x}) - \boldsymbol{f}(\hat{\boldsymbol{x}}) = D_{\boldsymbol{x}} \boldsymbol{f}(\hat{\boldsymbol{x}}) \Delta\boldsymbol{x} + h.o.t.$$