Multivariate Calculus – A Brief Reminder

Purpose. The purpose of this document is to quickly refresh (presumably) known notions in multivariate differential calculus such as differentials, directional derivatives, the gradient and the Hessian. These notions will be used heavily in our course. Even though this quick reminder may seem redundant or trivial to most of you (I hope), I still suggest at least to skim through it, as it might present less common ways of interpretation of very familiar definitions and properties. And even if you discover nothing new in this document, it will at least be useful to introduce notation.

1 Notation

In our course, we will deal exclusively with real functions. A scalar function will be denoted as $f : \mathbb{R}^n \to \mathbb{R}$, $f(\mathbf{x})$, or simply f. A vector-valued function will be denoted in bold, as $\mathbf{f} : \mathbb{R}^n \to \mathbb{R}^m$, or component-wise as $\mathbf{f}(\mathbf{x}) = (f_1(\mathbf{x}), \ldots, f_m(\mathbf{x}))^{\mathrm{T}}$. A scalar function of a matrix variable, $f : \mathbb{R}^{m\times n} \to \mathbb{R}$, will be denoted as $f(\mathbf{A})$, and a matrix-valued function of a vector, $f : \mathbb{R}^n \to \mathbb{R}^{m\times k}$ as $\mathbf{F}(\mathbf{x})$. Derivatives of a scalar function of one variable will be denoted as $f'(x)$, $f''(x)$, etc. An n-times continuously differentiable function will be said $\mathcal{C}^n$ ($f \in \mathcal{C}^n$). In most cases, we will tacitly assume that a function is sufficiently smooth for at least the first-order derivative to exist.

2 First-order derivative of a function of one variable

Before proceeding to multivariate functions, let us remind ourselves a few basic notions of univariate calculus. A $\mathcal{C}^1$ function $f(x)$ can be approximated linearly around some point $x = x_0$ (Figure 1). Incrementing the argument by dx, the function itself changes by the amount that we denote by $\Delta f = f(x_0 + dx) - f(x_0)$, while the linear approximation changes by the amount denoted by df. For a sufficiently small dx (more formally, in the limit $|dx| \to 0$), it can be shown that $\Delta f = df + o(dx)$[1]. This means that for an infinitesimally small increment dx, the linear approximation of the function becomes exact. In this limit, df is called the *differential* of f, and the slope of the linear approximation, is called the

[1]The little-o notation means that there exists some function of dx, $o(dx)$, going faster to zero than dx (i.e., $\frac{o(dx)}{dx} \to 0$), but the exact form of this function is unimportant. On the other hand, the big-O notation, as in $O(dx^2)$, stands for some function that grows with the same rate as dx^2 (i.e., $\lim_{|dx|\to 0} \frac{dx^2}{O(dx^2)} < \infty$).

Figure 1: First-order derivative of a function of one variable.

first-order derivative of f, denoted $\frac{df}{dx} = f'(x_0)$. Another way to express this fact is through the first-order *Taylor expansion* of f around x_0:

$$f(x_0 + dx) = f(x_0) + f'(x_0)dx + O(dx^2),$$

which essentially says that a linear function whose value at x_0 matches that of $f(x_0)$, and whose slope matches that of f (expressed by $f'(x_0)$) approximates f around x_0 up to some second-order error.

3 Gradient

We can extend the previous discussion straightforwardly to the n-dimensional case. Let f now be a $\mathcal{C}^1$ function on $\mathbb{R}^n$. The surface the function creates in $\mathbb{R}^{n+1}$ can be approximated by an n-dimensional tangent plane (the multidimensional analog of linear approximation). Fixing a point $\mathbf{x}_0$ and making a small step $\mathbf{dx} = (dx_1, \ldots, dx_n)^{\mathrm{T}}$ (note that now $\mathbf{dx}$ is a vector), it can be shown that the change in the value of the linear approximation is given by

$$df = \frac{\partial f}{\partial x_1}dx_1 + \cdots + \frac{\partial f}{\partial x_n}dx_n,$$

where $\frac{\partial f}{\partial x_i}$ denotes the *partial derivative* of f at $\mathbf{x}_0$. The latter formula is usually known as the *total differential*. Arranging the partial derivatives into a vector $\mathbf{g} = \left(\frac{\partial f}{\partial x_1}, \ldots, \frac{\partial f}{\partial x_n}\right)^{\mathrm{T}}$, the total differential can be expressed as the inner product $df = \langle \mathbf{g}, \mathbf{dx} \rangle$. The object $\mathbf{g}$ appearing in the inner product is called the *gradient* of f at point $\mathbf{x}_0$, and will be denoted by $\nabla f(\mathbf{x}_0)$ (the symbol ∇, graphically a rotated capital Delta, is pronounced "nabla", from the grecized Hebrew "nevel" for "harp"; ∇ is sometimes called the *del operator*). While we can simply define the gradient as the vector of partial derivatives, we will see that the definition through the inner product can often be more useful.

4 Directional derivative

In this course, we will often encounter situations where we are interested in the behavior of a function along a line (formally, we say that $f(\mathbf{x})$ is restricted to the one-dimensional linear subspace $\mathcal{L} = \{\mathbf{x}_0 + \alpha\mathbf{r} : \alpha \in \mathbb{R}\}$, where $\mathbf{x}_0$ is some fixed point, and $\mathbf{r}$ is a fixed direction). Let use define a new function of a single variable α, $\varphi(\alpha) = f(\mathbf{x}_0 + \alpha\mathbf{r})$. Note that we can find the first-order derivative of φ, arriving at the following important notion:

Definition. $f'_{\mathbf{r}}(\mathbf{x}_0) = \dfrac{d}{d\alpha} f(\mathbf{x}_0 + \alpha\mathbf{r})\Big|_{\alpha=0} = \varphi'(0)$ *is called the* directional derivative *of* f *at* $\mathbf{x}_0$ *in the direction* $\mathbf{r}$.

The same way a derivative measures the rate of change of a function, a directional derivative measures the rate of change of a multivariate function when we make a small step in a particular direction.

Denoting $\mathbf{g} = \nabla f(\mathbf{x}_0)$ and using our definition of the gradient as the inner product, we can write

$$d\varphi = df = \mathbf{g}^{\mathrm{T}}\mathbf{dx} = \mathbf{g}^{\mathrm{T}}(d\alpha\mathbf{r}) = d\alpha(\mathbf{g}^{\mathrm{T}}\mathbf{r}).$$

Identifying in the latter quantity an inner product of $d\alpha$ with the scalar $\mathbf{g}^{\mathrm{T}}\mathbf{r}$, we can say that $\mathbf{g}^{\mathrm{T}}\mathbf{r}$ is the gradient of $\varphi(\alpha)$ at $\alpha = 0$, which coincides with the first-order derivative, $\varphi'(0) = \mathbf{g}^{\mathrm{T}}\mathbf{r}$, as φ is a function of a single variable. We can summarize this result as the following

Property. *The directional derivative of* f *at* $\mathbf{x}_0$ *in the direction* $\mathbf{r}$ *is obtained by projecting the gradient at* $\mathbf{x}_0$ *onto the direction* $\mathbf{r}$, $f'_{\mathbf{r}} = \mathbf{r}^{\mathrm{T}}\nabla f(\mathbf{x}_0)$.

5 Hessian

In the case of a function of a single variable, we saw that the differential of f was given by $df = f'(x)dx$. However, the first-order derivative $f'(x)$ is also a function of x, and we can again express its differential as $df' = f''(x)dx$, where $f''(x)$ denotes the second-order derivative. This notion can be extended to the multivariate case. Recall our definition of the gradient through the inner product,

$$df = \mathbf{g}^{\mathrm{T}}\mathbf{dx}.$$

Thinking of the gradient as of a vector-valued function on $\mathbb{R}^n$, $\mathbf{g}(\mathbf{x}) = (g_1(\mathbf{x}), \ldots, g_n(\mathbf{x}))^{\mathrm{T}}$, we can write

$$\begin{cases} dg_1 &= \mathbf{h}_1^{\mathrm{T}}\mathbf{dx} \\ \vdots & \vdots \\ dg_n &= \mathbf{h}_n^{\mathrm{T}}\mathbf{dx}, \end{cases}$$

with each $\mathbf{h}_i$ being the gradient of the i-th component of the gradient vector $\mathbf{g}$,

$$\mathbf{h}_i = \left(\frac{\partial g_i}{\partial x_1}, \ldots, \frac{\partial g_i}{\partial x_n}\right)^{\mathrm{T}} = \left(\frac{\partial^2 f}{\partial x_1 \partial x_i}, \ldots, \frac{\partial^2 g_i}{\partial x_n \partial x_i}\right)^{\mathrm{T}}.$$

Denoting by $\mathbf{H} = (\mathbf{h}_1, \ldots, \mathbf{h}_n)$, we can write compactly $\mathbf{dg} = \mathbf{H}^{\mathrm{T}}\mathbf{dx}$. The $n \times n$ matrix $\mathbf{H}$ containing all the second-order partial derivatives of f as its elements is called the *Hessian* of

f at point $\mathbf{x}$, and is also denoted[2] as $\nabla^2 f(\mathbf{x})$. We tacitly assumed that f is $\mathcal{C}^2$ in order for the second-order derivatives to exist. A nice property of $\mathcal{C}^2$ functions is that partial derivation is commutative, meaning that the order of taking second-order partial derivatives can be interchanged: $h_{ij} = \dfrac{\partial^2 f}{\partial x_i \partial x_j} = \dfrac{\partial^2 f}{\partial x_j \partial x_i} = h_{ji}$. Algebraically, this implies that the Hessian matrix is symmetric, and we can write

$$\mathbf{dg} = \mathbf{H}\mathbf{dx}.$$

6 Second-order directional derivative

Recall that we have previously considered the restriction of a multivariate function f to a line, $\varphi(\alpha) = f(\mathbf{x}_0 + \alpha\mathbf{r})$. This gave rise to the first-order directional derivative $f_{\mathbf{r}}(\mathbf{x}_0) = \varphi'(0)$. In a similar way, we define the *second-order directional derivative* at $\mathbf{x}_0$ in the direction $\mathbf{r}$ as

$$f''_{\mathbf{rr}}(\mathbf{x}_0) \;=\; \varphi''(0) = \frac{d^2}{d\alpha^2} f(\mathbf{x}_0 + \alpha\mathbf{r})\bigg|_{\alpha=0} = \frac{d}{d\alpha} f'_{\mathbf{r}}(\mathbf{x}_0 + \alpha\mathbf{r})\bigg|_{\alpha=0}.$$

Considering $f'_{\mathbf{r}}(\mathbf{x}) = \mathbf{r}^{\mathrm{T}}\mathbf{g}(\mathbf{x})$ as a function of $\mathbf{x}$, we can write its differential as

$$df'_{\mathbf{r}} = \mathbf{r}^{\mathrm{T}}\mathbf{dg} = \mathbf{r}^{\mathrm{T}} bbH(\mathbf{x}_0)\mathbf{dx} = \mathbf{r}^{\mathrm{T}}\mathbf{H}(\mathbf{x}_0)\mathbf{r}d\alpha,$$

from where

$$f''_{\mathbf{rr}} = \mathbf{r}^{\mathrm{T}}\mathbf{H}\mathbf{r}.$$

In other words, in order to get the second-order directional derivative in the direction $\mathbf{r}$, one has to evaluate the quadratic form $\mathbf{r}^{\mathrm{T}}\mathbf{H}\mathbf{r}$.

7 Derivatives of linear and quadratic functions

Let $\mathbf{y} = \mathbf{A}\mathbf{x}$ be a general linear operator defined by an $m \times n$ matrix. Its differential is given straightforwardly by

$$\mathbf{dy} = \mathbf{A}(\mathbf{x} + \mathbf{dx}) - \mathbf{A}\mathbf{x} = \mathbf{A}\mathbf{dx}.$$

Using this result, we will do a small exercise deriving gradients and Hessians of linear and quadratic functions. As we will see, it is often convenient to start with evaluating the differential of a function.

---

[2] Some people find helpful the following abuse of notation: Thinking of the gradient of f as of a differential operator of the form "$\nabla = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{pmatrix}$" applied to f, the Hessian can be expressed by applying the operator "$\nabla^2 = \nabla\nabla^{\mathrm{T}} = \begin{pmatrix} \frac{\partial}{\partial x_1} \\ \vdots \\ \frac{\partial}{\partial x_n} \end{pmatrix} \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n} \right) = \begin{pmatrix} \frac{\partial^2}{\partial x_1 \partial x_1} & \cdots & \frac{\partial^2}{\partial x_1 \partial x_n} \\ \vdots & \ddots & \vdots \\ \frac{\partial^2}{\partial x_n \partial x_1} & \cdots & \frac{\partial^2}{\partial x_n \partial x_n} \end{pmatrix}$".

Our first example is a linear function of the form $f(\mathbf{x}) = \mathbf{b}^{\mathrm{T}}\mathbf{x}$, where $\mathbf{b}$ is a constant vector. Note that this function is a particular case of the previous result (with $\mathbf{A} = \mathbf{b}^{\mathrm{T}}$), and we can write $df = \mathbf{b}^{\mathrm{T}}\mathbf{dx}$. Comparing this to the general definition of the gradient, $df = \mathbf{g}^{\mathrm{T}}(\mathbf{x})\mathbf{dx}$, we deduce that the gradient of f is given by $\nabla f(\mathbf{x}) = \mathbf{b}$. Note that the gradient of a linear function is constant – this generalizes the case of a linear function of one variable, $f(x) = bx$, which has a constant derivative $f'(x) = b$.

Our second example is a quadratic function of the form $f(\mathbf{x}) = \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x}$, where $\mathbf{A}$ is an $n \times n$ matrix. We again compute the differential by definition,

$$\begin{aligned} df &= f(\mathbf{x}+\mathbf{dx}) - f(\mathbf{x}) = (\mathbf{x}+\mathbf{dx})^{\mathrm{T}}\mathbf{A}(\mathbf{x}+\mathbf{dx}) - \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} \\ &= \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} + \mathbf{dx}^{\mathrm{T}}\mathbf{A}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{dx} + \mathbf{dx}^{\mathrm{T}}\mathbf{A}\mathbf{dx} - \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{x} \\ &= \mathbf{dx}^{\mathrm{T}}\mathbf{A}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{dx} + \mathbf{dx}^{\mathrm{T}}\mathbf{A}\mathbf{dx}. \end{aligned}$$

Note that in the limit $\|\mathbf{dx}\| \to 0$, the third term (quadratic in $\|\mathbf{dx}\|$) goes to zero much faster than the first two terms (linear in $\mathbf{dx}$), and can be therefore neglected[3], leading to

$$df = \mathbf{dx}^{\mathrm{T}}\mathbf{A}\mathbf{x} + \mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{dx} = \mathbf{dx}^{\mathrm{T}}\mathbf{A}\mathbf{x} + (\mathbf{x}^{\mathrm{T}}\mathbf{A}\mathbf{dx})^{\mathrm{T}} = \mathbf{dx}^{\mathrm{T}}(\mathbf{A}^{\mathrm{T}} + \mathbf{A})\mathbf{x}.$$

Again, recognizing in the latter expression an inner product with $\mathbf{dx}$, we conclude that $\nabla f(\mathbf{x}) = (\mathbf{A}^{\mathrm{T}} + \mathbf{A})\mathbf{x}$. For a symmetric $\mathbf{A}$, the latter simplifies to $\nabla f(\mathbf{x}) = 2\mathbf{A}\mathbf{x}$. Note that the gradient of a quadratic function is a linear function; furthermore, the latter expression generalizes the univariate quadratic function $f(x) = ax^2$, whose first-order derivative $f'(x) = 2ax$ is linear.

Since the gradient $\mathbf{g}(\mathbf{x}) = (\mathbf{A}^{\mathrm{T}} + \mathbf{A})\mathbf{x}$ of the quadratic function is linear, its differential is immediately given by $\mathbf{dg} = (\mathbf{A}^{\mathrm{T}} + \mathbf{A})\mathbf{dx}$, from where we conclude that the Hessian of f is $\mathbf{H}(\mathbf{x}) = \mathbf{A}^{\mathrm{T}} + \mathbf{A}$ (or $2\mathbf{A}$ in the symmetric case). Note that the Hessian of a quadratic function is constant, which coincides with the univariate case $f''(x) = 2a$.

In the sequel, we will see more complicated examples of gradients and Hessians. For a comprehensive reference on derivatives of matrix and vector expressions, the Matrix Cookbook[4] is highly advisable.

8 Multivariate Taylor expansion

We have seen the Taylor expansion of a function of one variable as a way to obtain a linear approximation. This construction can be generalized to the multivariate case, as we show here, limiting the expansion to second order.

[3]This "explanation" can be written rigorously using limits. Another way of getting the same result is the well-known rule of "differential of a product", $d(fg) = df\,g + f\,dg$, which can be generalized to the multivariate case as follows: Let h be a scalar function given as the inner product of two vector-valued function, $h(\mathbf{x}) = \mathbf{f}^{\mathrm{T}}(\mathbf{x})\mathbf{g}(\mathbf{x})$. Then, $dh = \mathbf{df}^{\mathrm{T}}\mathbf{g} + \mathbf{f}^{\mathrm{T}}\mathbf{dg}$.

[4]`http://www2.imm.dtu.dk/pubdb/views/edoc_download.php/3274/pdf/imm3274.pdf`

Theorem (Second-order Taylor expansion)**.** *Let f be a $\mathcal{C}^2$ function on $\mathbb{R}^n$, $\mathbf{x}$ some point, and $\mathbf{r}$ a sufficient small vector. Then,*

$$f(\mathbf{x}+\mathbf{r}) = f(\mathbf{x}) + \mathbf{g}^{\mathrm{T}}(\mathbf{x})\mathbf{r} + \frac{1}{2}\mathbf{r}^{\mathrm{T}}\mathbf{H}(\mathbf{x})\mathbf{r} + O(\|\mathbf{r}\|^3).$$

The theorem say that up to a third-order error term, the function can be approximated around $\mathbf{x}$ by a quadratic function $q(\mathbf{r}) = f + \mathbf{g}^{\mathrm{T}}\mathbf{r} + \frac{1}{2}\mathbf{r}^{\mathrm{T}}\mathbf{H}\mathbf{r}$ (note that the function is quadratic in $\mathbf{r}$, as $\mathbf{x}$ is constant, and so are $f = f(\mathbf{x})$, $\mathbf{g}$, and $\mathbf{H}$). Out of all possible quadratic approximations of f, the approximation described by $q(\mathbf{r}) \approx f(\mathbf{x}+\mathbf{r})$ is such that its value, slope, and curvature at $\mathbf{x}$ (equivalently, at $\mathbf{r} = \mathbf{0}$) match those of f. The latter geometric quantities are captured, respectively, by the values of the function, its gradient, and its Hessian; in order to match the value, slope, and curvature of f, q has to satisfy $q(\mathbf{0}) = f(\mathbf{x})$, $\nabla q(\mathbf{0}) = \nabla f(\mathbf{x})$, and $\nabla^2 q(\mathbf{0}) = \nabla^2 f(\mathbf{x})$ (note that the gradient and the Hessian of q are w.r.t $\mathbf{r}$, whereas the derivatives of f are w.r.t. $\mathbf{x}$). To see that the later equalities hold, we first observe that $q(\mathbf{0}) = f(\mathbf{x})$. Next, using the fact that $q(\mathbf{r})$ is quadratic, its gradient and Hessian (w.r.t. $\mathbf{r}$) are given by $\nabla q(\mathbf{r}) = \mathbf{g} + \mathbf{H}\mathbf{r}$ and $\nabla^2 q(\mathbf{r}) = \mathbf{H}\mathbf{r}$. Substituting $\mathbf{r} = \mathbf{0}$ yields $\nabla q(\mathbf{0}) = \mathbf{g}$ and $\nabla^2 q(\mathbf{r}) = \mathbf{H}$.

9 Gradient of a function of a matrix

The notion of gradient can be generalized to functions of matrices. Let $f : \mathbb{R}^{m\times n} \to \mathbb{R}$ be such function evaluated at some $\mathbf{X}$. We can think of an equivalent function on $\mathbb{R}^{mn}$ evaluated at $\mathbf{x} = \mathrm{vec}(\mathbf{X})$, for which the gradient is defined simply as the mn-dimensional vector of all partial derivatives. We can therefore think of the gradient of $f(\mathbf{X})$ at $\mathbf{X}$ as of the $m \times n$ matrix

$$\mathbf{G}(\mathbf{X}) = \begin{pmatrix} \dfrac{\partial f}{\partial x_{11}} & \cdots & \dfrac{\partial f}{\partial x_{1n}} \\ \vdots & \ddots & \vdots \\ \dfrac{\partial f}{\partial x_{m1}} & \cdots & \dfrac{\partial f}{\partial x_{mn}} \end{pmatrix}.$$

Previously, we have seen that an "external" definition of the gradient through an inner product is often more useful. Such a definition is also valid for matrix arguments. Recall our definition of the standard inner product on the space of $m \times n$ matrices as

$$\langle \mathbf{A}, \mathbf{B} \rangle = \sum_{ij} a_{ij} b_{ij} = \mathrm{tr}(\mathbf{A}^{\mathrm{T}}\mathbf{B}),$$

for $\mathbf{A}, \mathbf{B} \in \mathbb{R}^{m\times n}$. Using the total differential formula yields

$$df = \sum_{ij} \frac{\partial f}{\partial x_{ij}} dx_{ij} = \langle \mathbf{G}, \mathbf{dX} \rangle,$$

where $\mathbf{dX}$ is now an $m \times n$ matrix. The matrix $\mathbf{G}$ appearing in the above identity can be *defined* as the gradient of f.

10 Gradient of a nonlinear function

We finish this brief introduction by deriving the gradient of a more complicated function of the form

$$f(\mathbf{X}) = \mathbf{c}^{\mathrm{T}}\varphi(\mathbf{X}^{\mathrm{T}}\mathbf{a} + \mathbf{b}),$$

where $\mathbf{X} \in \mathbb{R}^{m\times n}$, $\mathbf{a} \in \mathbb{R}^n$, $\mathbf{b}, \mathbf{c} \in \mathbb{R}^m$, and φ is a $\mathcal{C}^1$ scalar applied element-wise. We will encounter such functions during the course when dealing with nonlinear regression and classification applications. In the machine learning, functions of this form constitute building blocks of more complicated functions called artificial neural networks.

As before, we proceed by computing differentials and using the chain rule. Denoting $\mathbf{u} = \mathbf{X}^{\mathrm{T}}\mathbf{a} + \mathbf{b}$, we have

$$\varphi(\mathbf{u}) = \begin{pmatrix} \varphi(u_1) \\ \vdots \\ \varphi(u_m) \end{pmatrix}.$$

Since φ is applied element-wise to $\mathbf{u}$, the differential of $\boldsymbol{\varphi} = \varphi(\mathbf{u})$ is given by

$$\mathbf{d}\boldsymbol{\varphi} = \begin{pmatrix} \varphi'(u_1)du_1 \\ \vdots \\ \varphi'(u_m)du_m \end{pmatrix} = \underbrace{\begin{pmatrix} \varphi'(u_1) & & \\ & \ddots & \\ & & \varphi'(u_m) \end{pmatrix}}_{\boldsymbol{\Phi}'} \mathbf{du} = \boldsymbol{\Phi}'\mathbf{du}.$$

Next, we consider the function $\mathbf{u}(\mathbf{X}) = \mathbf{X}^{\mathrm{T}}\mathbf{a} + \mathbf{b}$; since it is linear in $\mathbf{X}$, its differential is given by $\mathbf{du} = \mathbf{dX}^{\mathrm{T}}\mathbf{a}$. Finally, we consider the function $f(\boldsymbol{\varphi}) = \mathbf{c}^{\mathrm{T}}\boldsymbol{\varphi}$, which is linear in $\boldsymbol{\varphi}$ and has the differential $df = \mathbf{c}^{\mathrm{T}}\mathbf{d}\boldsymbol{\varphi}$.

Combining these results and using simple properties of the matrix trace yields

$$\begin{aligned} df &= \mathbf{c}^{\mathrm{T}}\mathbf{d}\boldsymbol{\varphi} = \mathbf{c}^{\mathrm{T}}\boldsymbol{\Phi}'\mathbf{du} = \mathbf{c}^{\mathrm{T}}\boldsymbol{\Phi}'\mathbf{dX}^{\mathrm{T}}\mathbf{a} \\ &= \mathrm{tr}\left(\mathbf{c}^{\mathrm{T}}\boldsymbol{\Phi}'\mathbf{dX}^{\mathrm{T}}\mathbf{a}\right) = \mathrm{tr}\left(\mathbf{dX}^{\mathrm{T}}\mathbf{ac}^{\mathrm{T}}\boldsymbol{\Phi}'\right) \\ &= \langle \mathbf{dX}, \mathbf{ac}^{\mathrm{T}}\boldsymbol{\Phi}' \rangle. \end{aligned}$$

In the latter expression, we recognize in the second argument of the inner product the gradient of f w.r.t. $\mathbf{X}$,

$$\nabla f(\mathbf{X}) = \mathbf{ac}^{\mathrm{T}}\boldsymbol{\Phi}'.$$