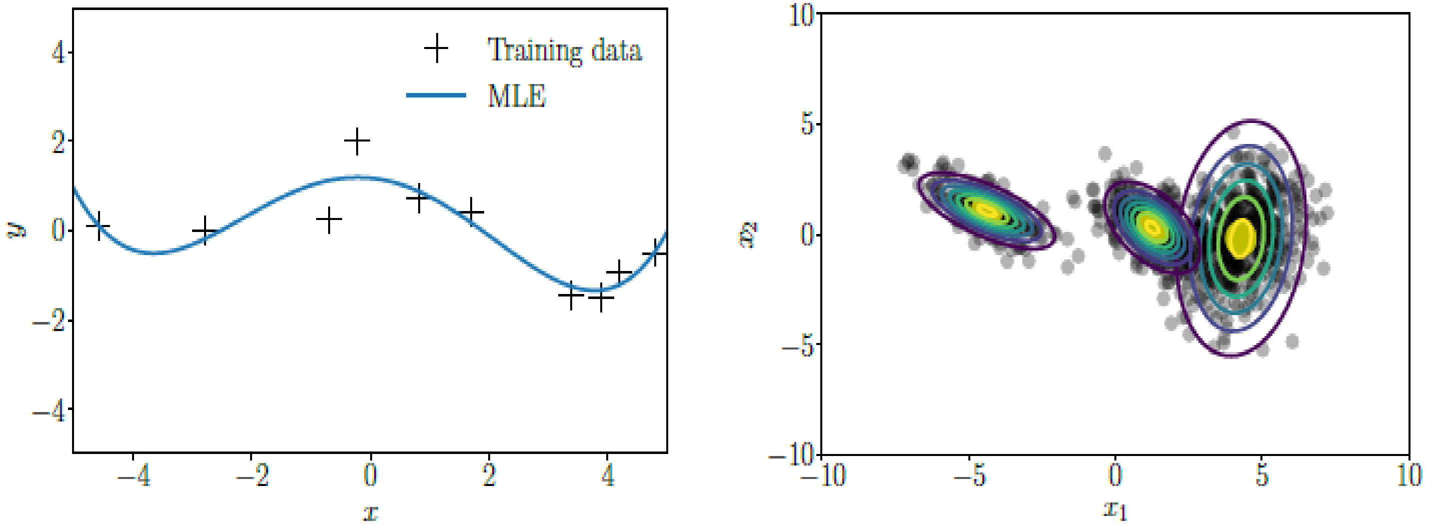
Vector Calculus

- Many algorithms in machine learning optimize an objective function with respect to a set of desired model parameters that control how well a model explains the data:
- Finding good parameters can be phrased as an optimization problem.
- Examples include: (i) linear regression, where we look at curve-fitting problems and optimize linear weight parameters to maximize the likelihood.
- (ii) neural-network auto-encoders for dimensionality reduction and data compression, where the parameters are the weights and biases of each layer, and where we minimize a reconstruction error by repeated application of the chain rule; and
- (iii) Gaussian mixture models for modeling data distributions, where we optimize the location and shape parameters of each mixture component to maximize the likelihood of the model.

These problems, which we typically solve by using optimization algorithms that exploit gradient information

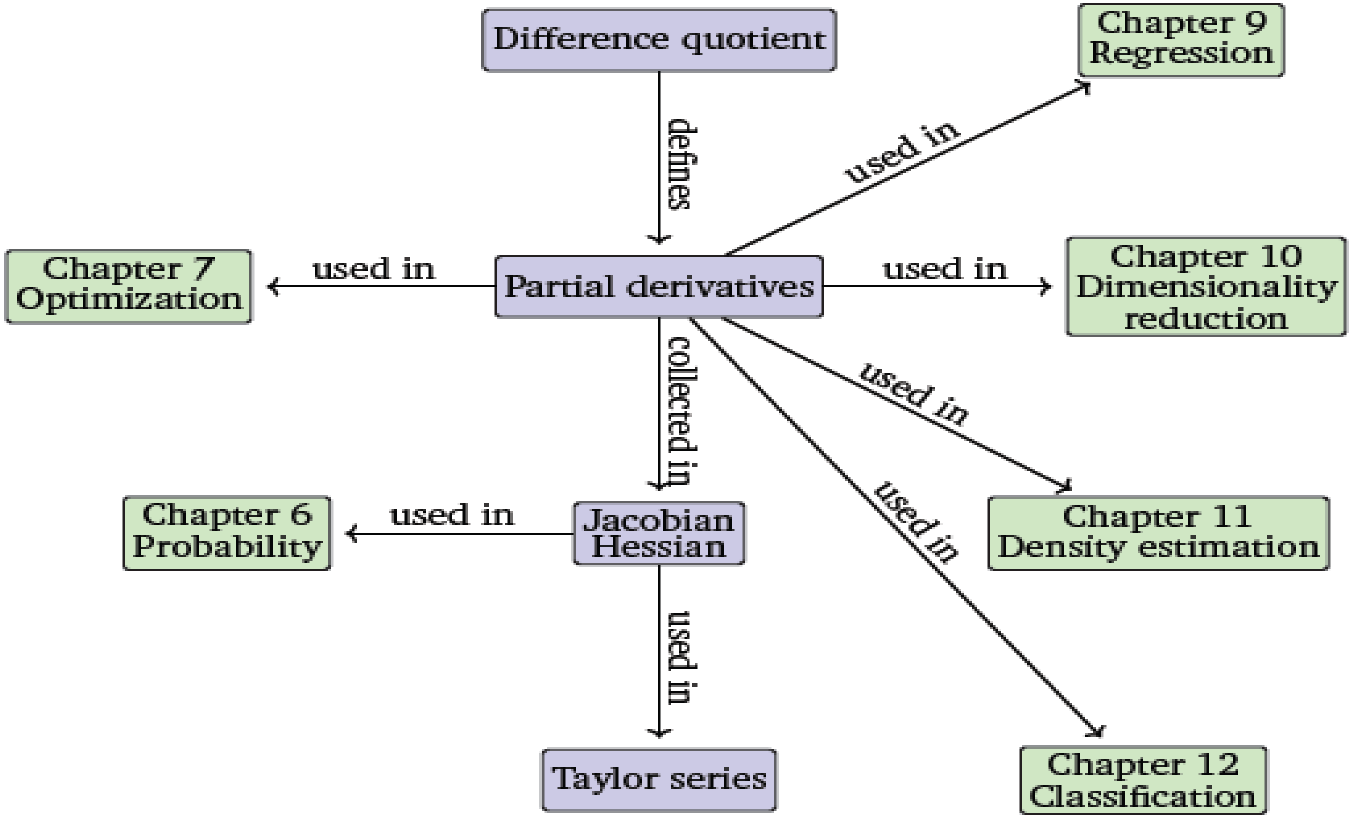


(a) Regression problem: Find parameters, such that the curve explains the observations (crosses) well.

(b) Density estimation with a Gaussian mixture model: Find means and covariances, such that the data (dots) can be explained well.

- Central to this vector calculus is the concept of a function.
  - A function f is a quantity that relates two quantities to each other.
  - These quantities are typically inputs x ϵ R^D and targets (function values) f(x), which we assume are real-valued if not stated otherwise.
  - Here R^D is the *domain of f, and the function values f(x) are the image/codomain of f.*

Use of Partial Differentiation



linear functions. We often write

$$f : \mathbb{R}^D \to \mathbb{R} \tag{5.1a}$$

$$\boldsymbol{x} \mapsto f(\boldsymbol{x}) \tag{5.1b}$$

to specify a function, where (**5.1a**) specifies that f is a mapping from $\mathbb{R}^D$ to $\mathbb{R}$ and (**5.1b**) specifies the explicit assignment of an input $\boldsymbol{x}$ to a function value $f(\boldsymbol{x})$. A function f assigns every input $\boldsymbol{x}$ exactly one function value $f(\boldsymbol{x})$.

Example 5.1

Recall the dot product as a special case of an inner product (Section **3.2**). In the previous notation, the function $f(\boldsymbol{x}) = \boldsymbol{x}^\top \boldsymbol{x}$, $\boldsymbol{x} \in \mathbb{R}^2$, would be specified as

$$f : \mathbb{R}^2 \to \mathbb{R} \tag{5.2a}$$

$$\boldsymbol{x} \mapsto x_1^2 + x_2^2 \,. \tag{5.2b}$$

Computation of Gradient

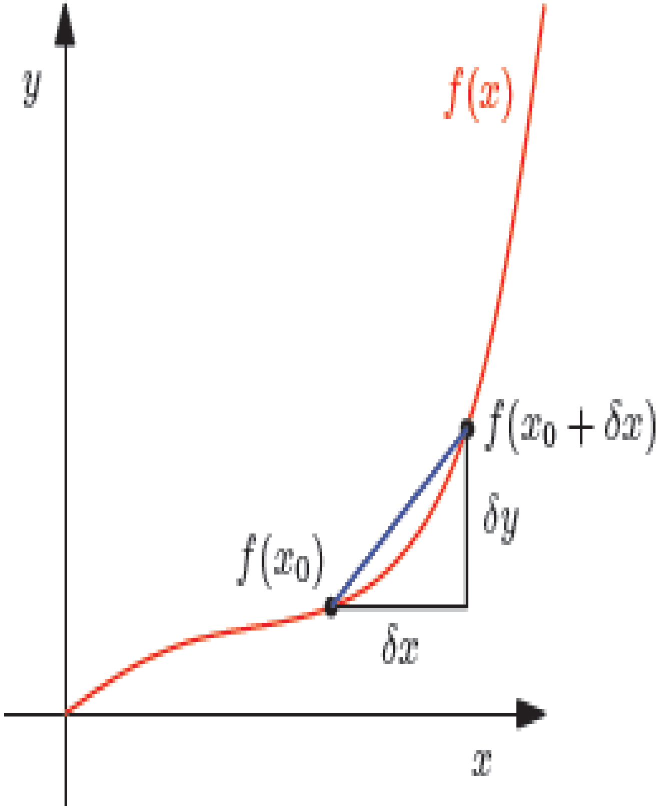
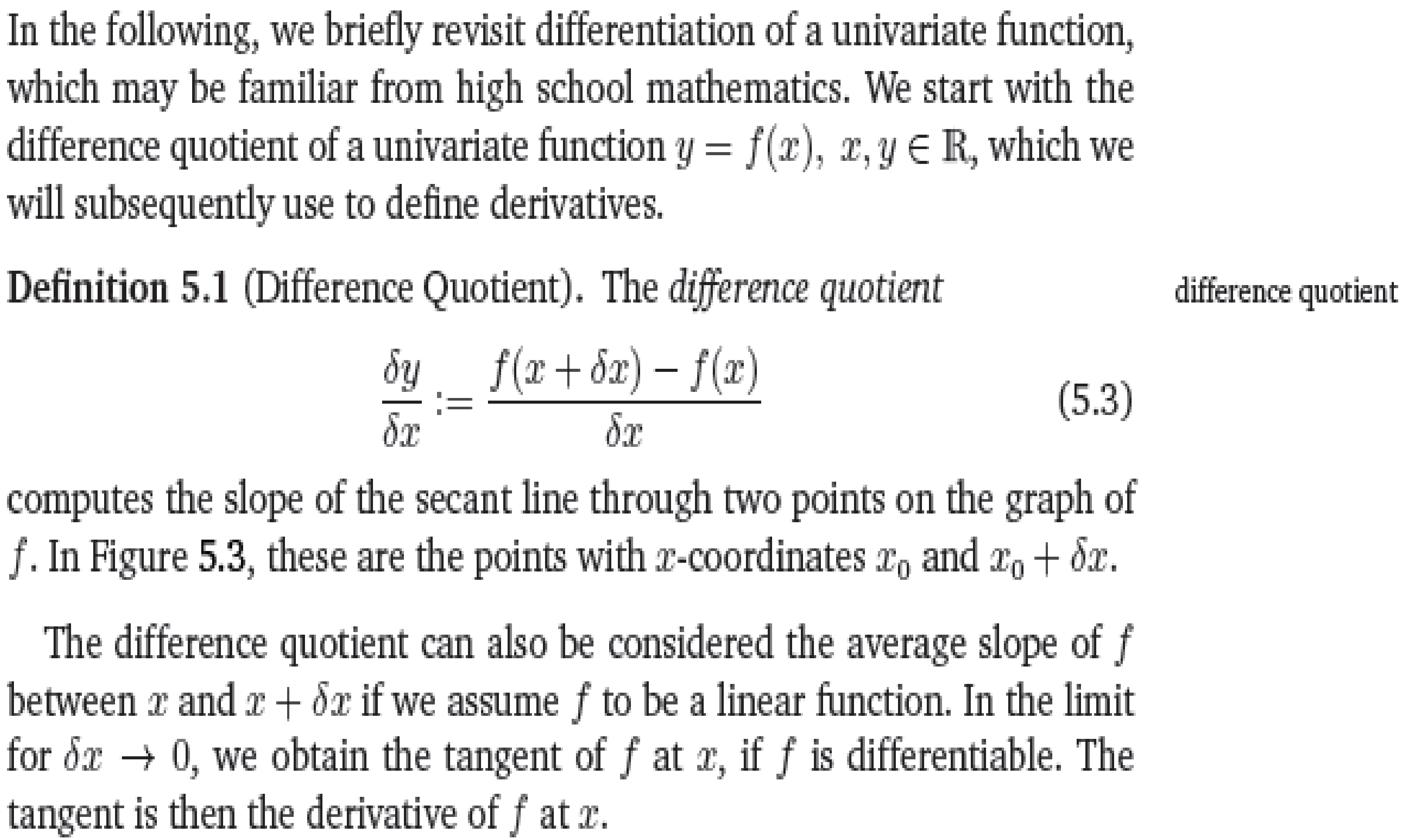


Figure 5.3 The average incline of a function f between x_0 and $x_0 + \delta x$ is the incline of the secant (blue) through $f(x_0)$ and $f(x_0 + \delta x)$ and given by $\delta y/\delta x$.

Differentiation of Univariate Functions



In the following, we briefly revisit differentiation of a univariate function, which may be familiar from high school mathematics. We start with the difference quotient of a univariate function $y = f(x),\ x, y \in \mathbb{R}$, which we will subsequently use to define derivatives.

Definition 5.1 (Difference Quotient). The *difference quotient*

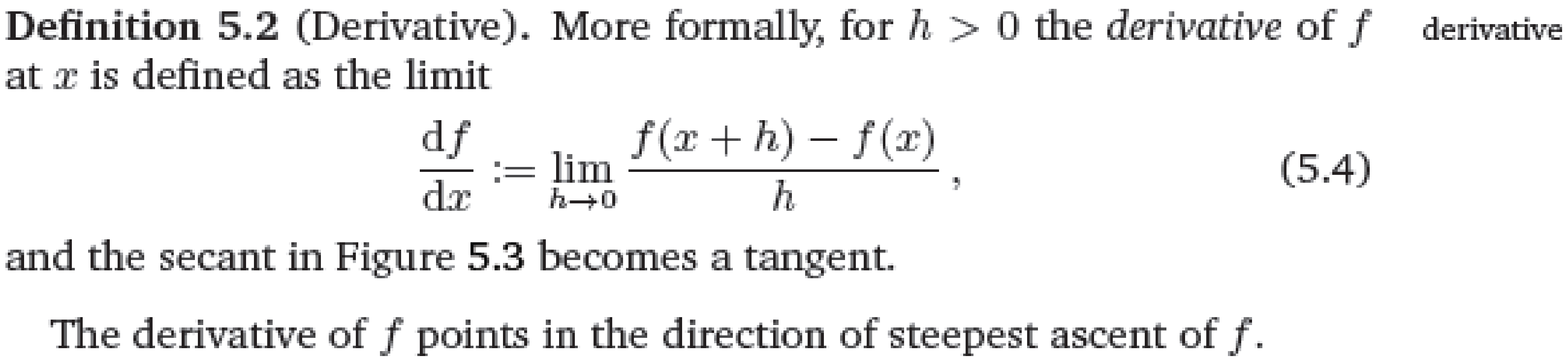
difference quotient

$$\frac{\delta y}{\delta x} := \frac{f(x + \delta x) - f(x)}{\delta x} \tag{5.3}$$

computes the slope of the secant line through two points on the graph of f. In Figure 5.3, these are the points with x-coordinates x_0 and $x_0 + \delta x$.

The difference quotient can also be considered the average slope of f between x and $x + \delta x$ if we assume f to be a linear function. In the limit for $\delta x \to 0$, we obtain the tangent of f at x, if f is differentiable. The tangent is then the derivative of f at x.

Derivative



Definition 5.2 (Derivative)**.** More formally, for $h > 0$ the *derivative* of f at x is defined as the limit

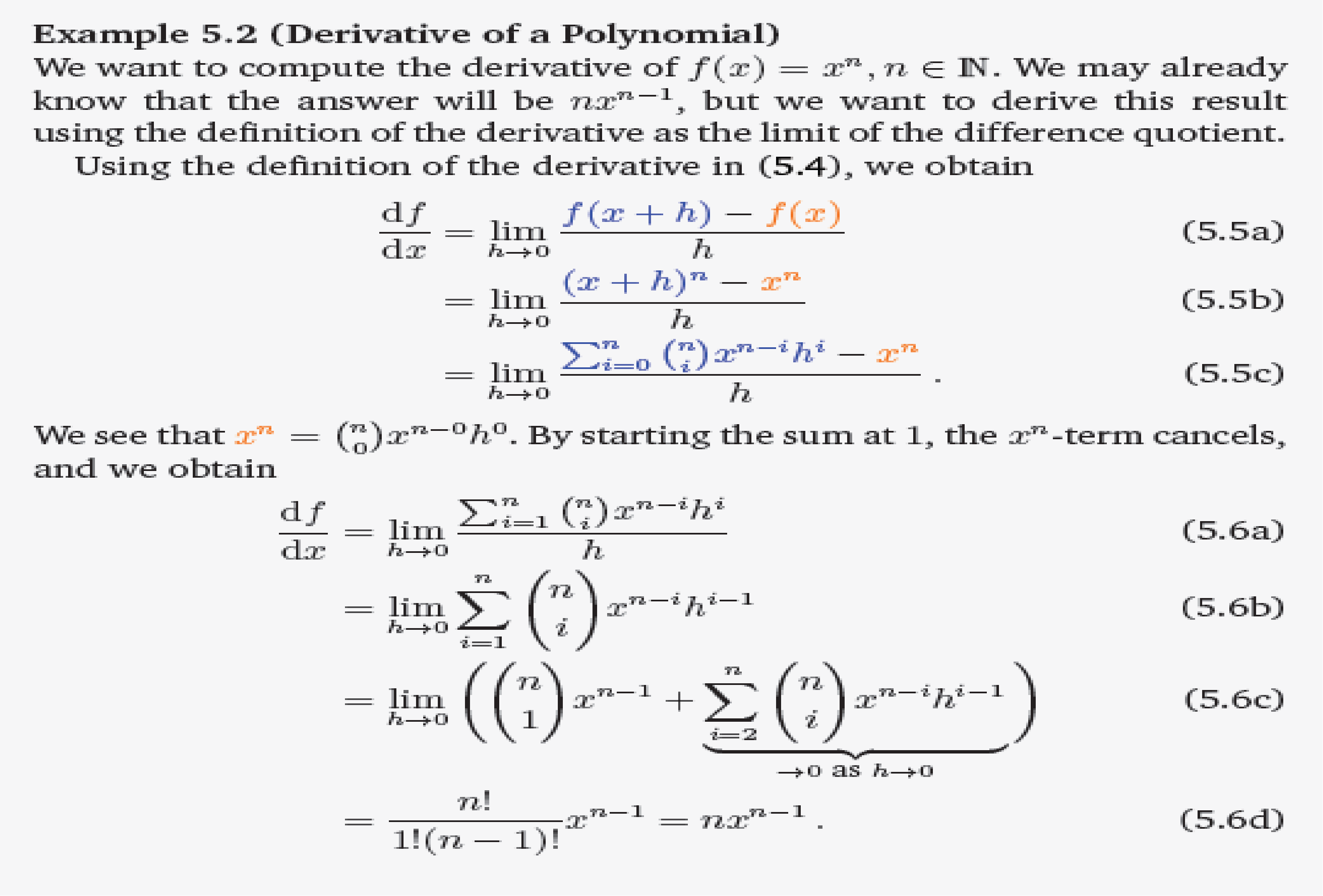
derivative

$$\frac{\mathrm{d}f}{\mathrm{d}x} := \lim_{h \to 0} \frac{f(x+h) - f(x)}{h}\,, \tag{5.4}$$

and the secant in Figure **5.3** becomes a tangent.

The derivative of f points in the direction of steepest ascent of f.

Deriavative Examples



Example 5.2 (Derivative of a Polynomial)
We want to compute the derivative of $f(x) = x^n, n \in \mathbb{N}$. We may already know that the answer will be nx^{n-1}, but we want to derive this result using the definition of the derivative as the limit of the difference quotient.

Using the definition of the derivative in (5.4), we obtain

$$\frac{\mathrm{d}f}{\mathrm{d}x} = \lim_{h\to 0} \frac{f(x+h) - f(x)}{h} \tag{5.5a}$$

$$= \lim_{h\to 0} \frac{(x+h)^n - x^n}{h} \tag{5.5b}$$

$$= \lim_{h\to 0} \frac{\sum_{i=0}^{n} \binom{n}{i} x^{n-i} h^i - x^n}{h}\,. \tag{5.5c}$$

We see that $x^n = \binom{n}{0} x^{n-0} h^0$. By starting the sum at 1, the x^n-term cancels, and we obtain

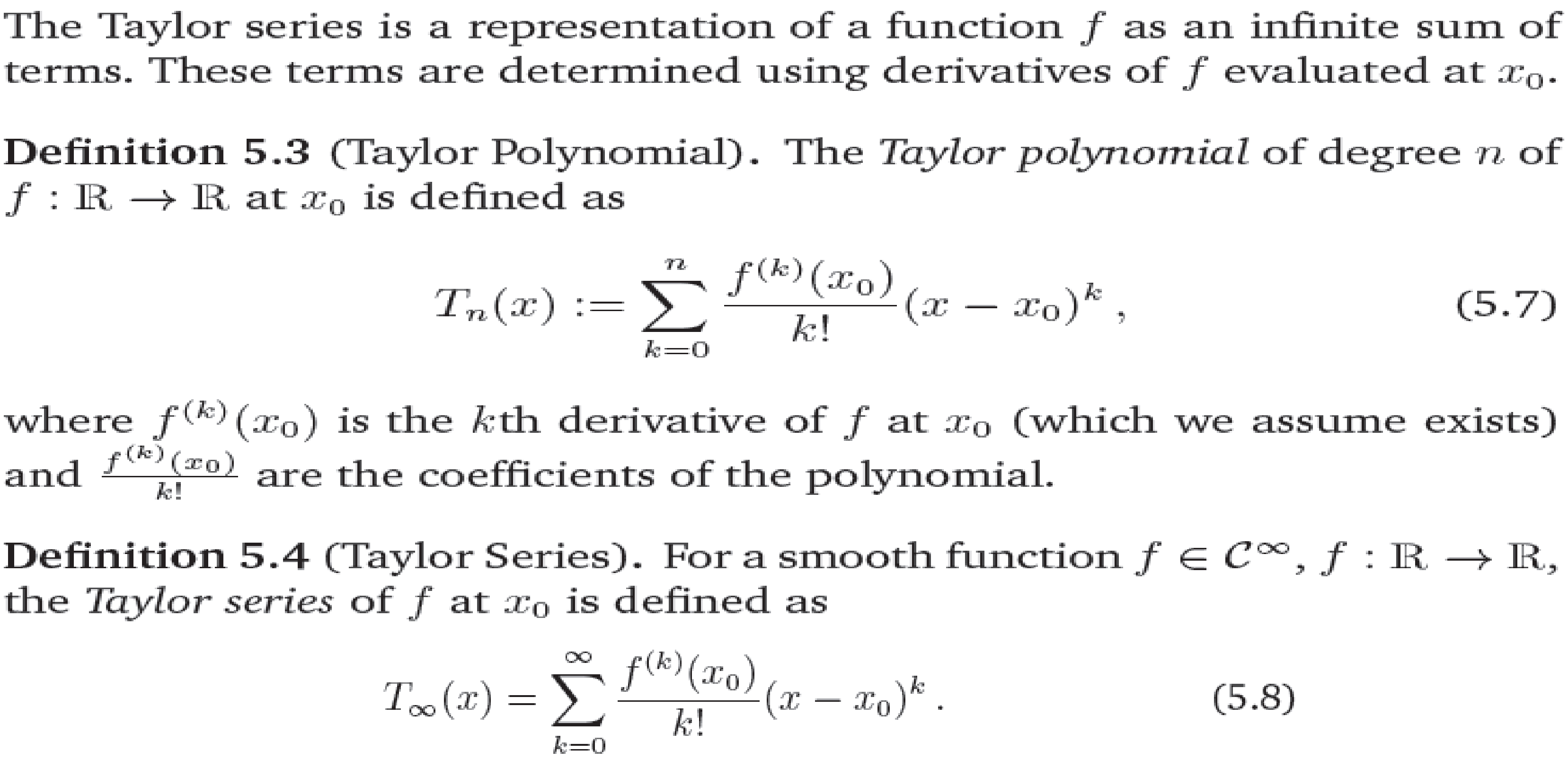
$$\frac{\mathrm{d}f}{\mathrm{d}x} = \lim_{h\to 0} \frac{\sum_{i=1}^{n} \binom{n}{i} x^{n-i} h^i}{h} \tag{5.6a}$$

$$= \lim_{h\to 0} \sum_{i=1}^{n} \binom{n}{i} x^{n-i} h^{i-1} \tag{5.6b}$$

$$= \lim_{h\to 0} \left(\binom{n}{1} x^{n-1} + \underbrace{\sum_{i=2}^{n} \binom{n}{i} x^{n-i} h^{i-1}}_{\to 0 \text{ as } h \to 0} \right) \tag{5.6c}$$

$$= \frac{n!}{1!(n-1)!} x^{n-1} = n x^{n-1}\,. \tag{5.6d}$$

Taylor Series



The Taylor series is a representation of a function f as an infinite sum of terms. These terms are determined using derivatives of f evaluated at x_0.

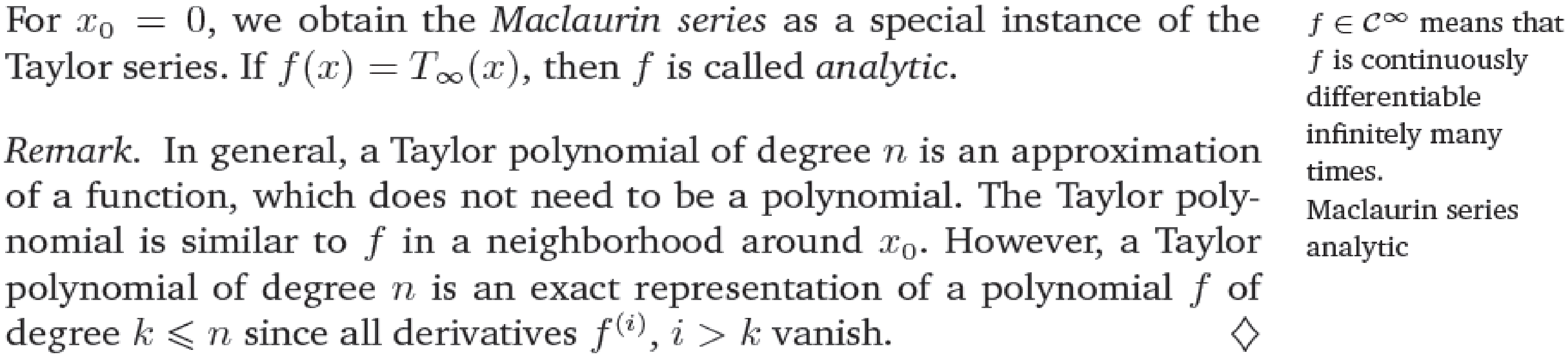
Definition 5.3 (Taylor Polynomial). The *Taylor polynomial* of degree n of $f : \mathbb{R} \to \mathbb{R}$ at x_0 is defined as

$$T_n(x) := \sum_{k=0}^{n} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \,, \tag{5.7}$$

where $f^{(k)}(x_0)$ is the kth derivative of f at x_0 (which we assume exists) and $\frac{f^{(k)}(x_0)}{k!}$ are the coefficients of the polynomial.

Definition 5.4 (Taylor Series). For a smooth function $f \in \mathcal{C}^\infty$, $f : \mathbb{R} \to \mathbb{R}$, the *Taylor series* of f at x_0 is defined as

$$T_\infty(x) = \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \,. \tag{5.8}$$



For $x_0 = 0$, we obtain the *Maclaurin series* as a special instance of the Taylor series. If $f(x) = T_\infty(x)$, then f is called *analytic*.

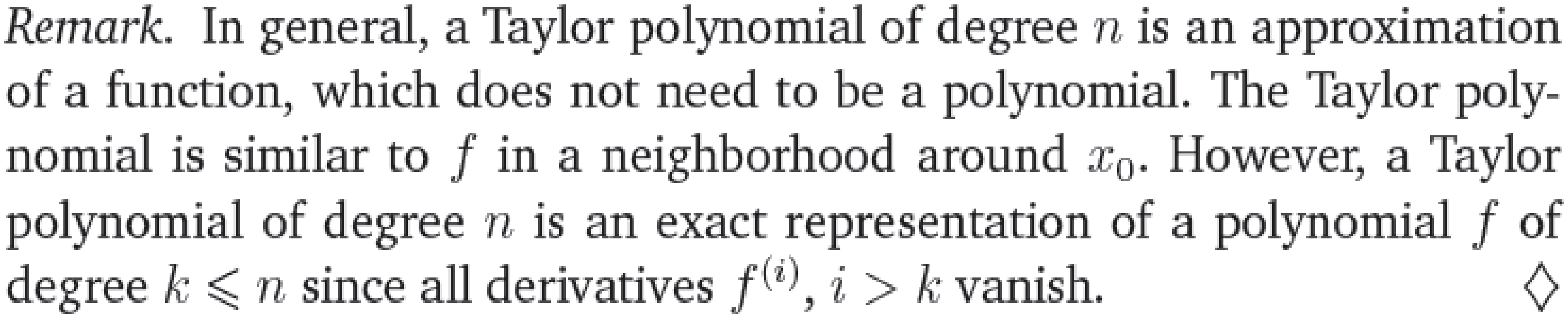
$f \in \mathcal{C}^\infty$ means that f is continuously differentiable infinitely many times.

Maclaurin series

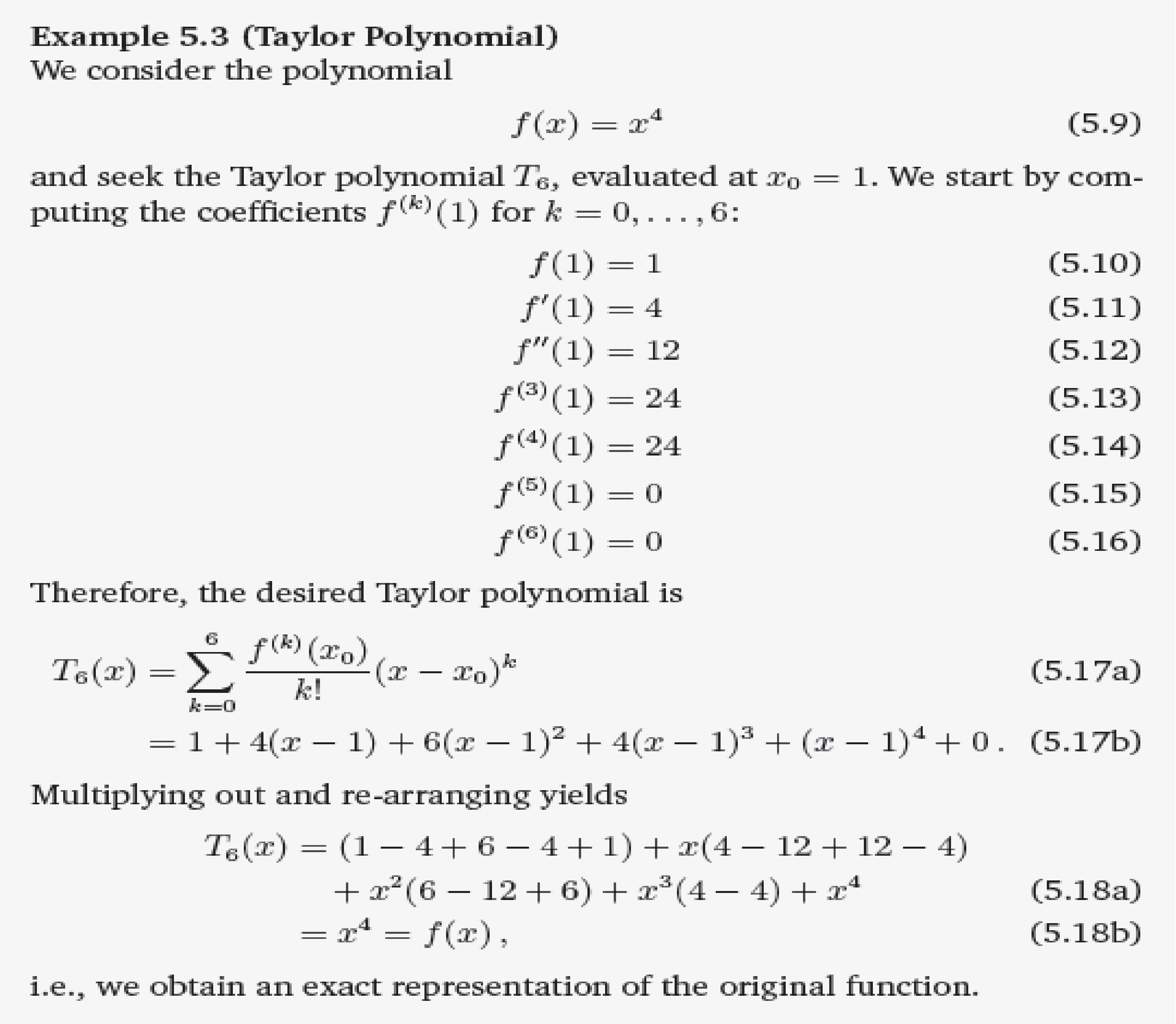
analytic

Remark. In general, a Taylor polynomial of degree n is an approximation of a function, which does not need to be a polynomial. The Taylor polynomial is similar to f in a neighborhood around x_0. However, a Taylor polynomial of degree n is an exact representation of a polynomial f of degree $k \leqslant n$ since all derivatives $f^{(i)}$, $i > k$ vanish. ♢

Taylor Series



Remark. In general, a Taylor polynomial of degree n is an approximation of a function, which does not need to be a polynomial. The Taylor polynomial is similar to f in a neighborhood around x_0. However, a Taylor polynomial of degree n is an exact representation of a polynomial f of degree $k \leqslant n$ since all derivatives $f^{(i)}$, $i > k$ vanish. $\diamondsuit$



Example 5.3 (Taylor Polynomial)
We consider the polynomial

$$f(x) = x^4 \tag{5.9}$$

and seek the Taylor polynomial T_6, evaluated at $x_0 = 1$. We start by computing the coefficients $f^{(k)}(1)$ for $k = 0, \ldots, 6$:

$$f(1) = 1 \tag{5.10}$$

$$f'(1) = 4 \tag{5.11}$$

$$f''(1) = 12 \tag{5.12}$$

$$f^{(3)}(1) = 24 \tag{5.13}$$

$$f^{(4)}(1) = 24 \tag{5.14}$$

$$f^{(5)}(1) = 0 \tag{5.15}$$

$$f^{(6)}(1) = 0 \tag{5.16}$$

Therefore, the desired Taylor polynomial is

$$T_6(x) = \sum_{k=0}^{6} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k \tag{5.17a}$$

$$= 1 + 4(x-1) + 6(x-1)^2 + 4(x-1)^3 + (x-1)^4 + 0\,. \tag{5.17b}$$

Multiplying out and re-arranging yields

$$\begin{aligned} T_6(x) &= (1 - 4 + 6 - 4 + 1) + x(4 - 12 + 12 - 4) \\ &\quad + x^2(6 - 12 + 6) + x^3(4 - 4) + x^4 \end{aligned} \tag{5.18a}$$

$$= x^4 = f(x)\,, \tag{5.18b}$$

i.e., we obtain an exact representation of the original function.

Taylor Polynomials

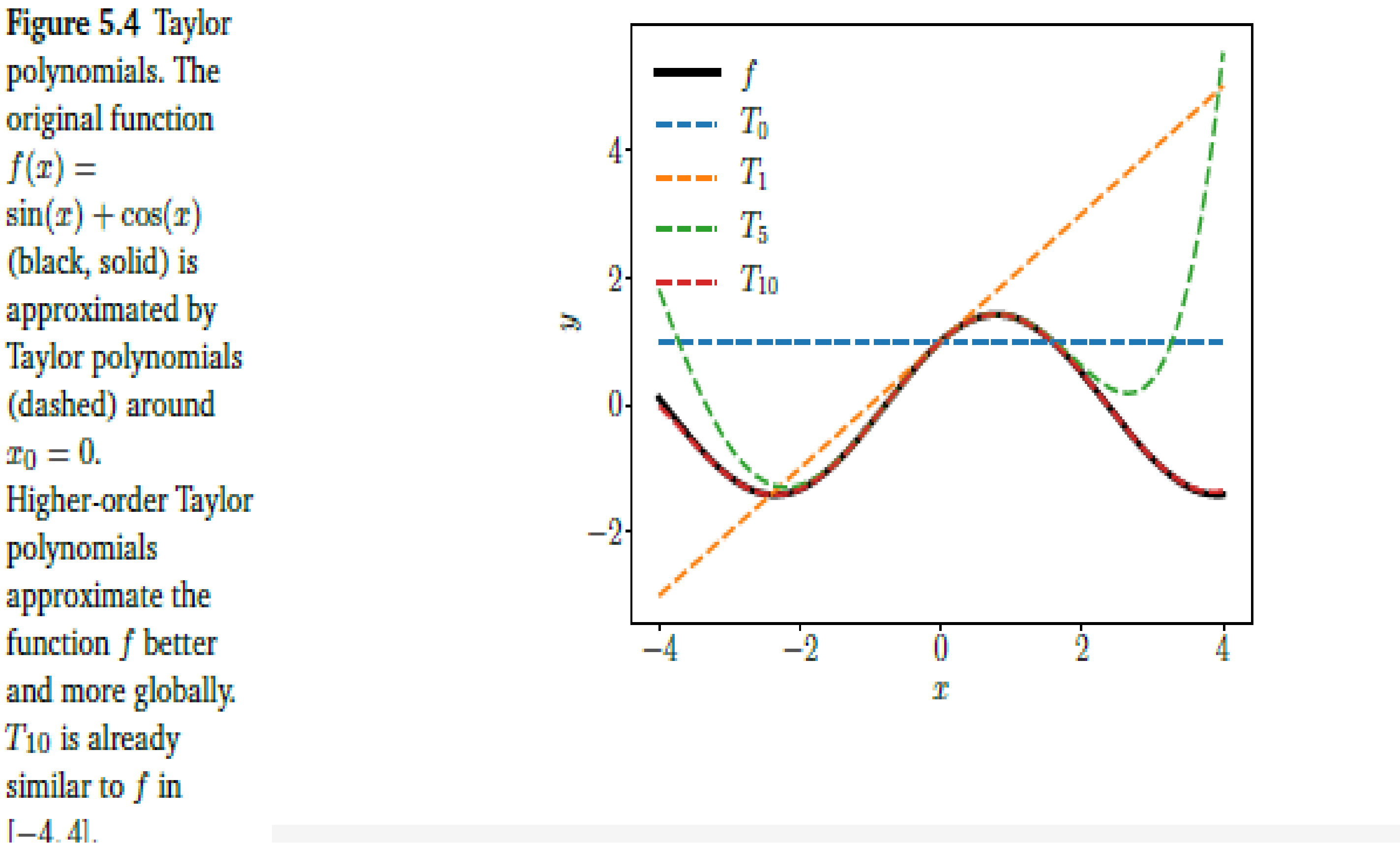
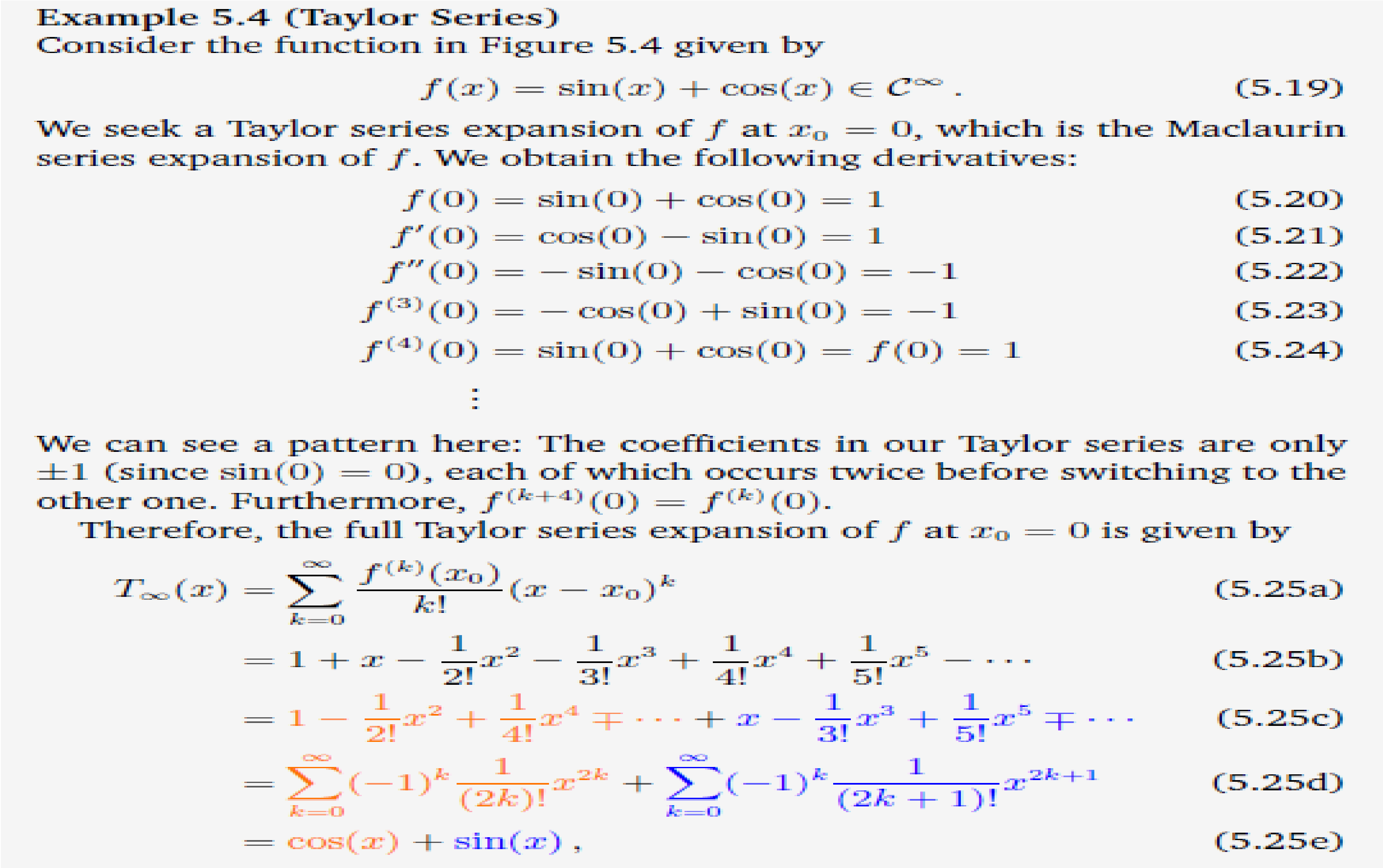


Figure 5.4 Taylor polynomials. The original function $f(x) = \sin(x) + \cos(x)$ (black, solid) is approximated by Taylor polynomials (dashed) around $x_0 = 0$. Higher-order Taylor polynomials approximate the function f better and more globally. T_{10} is already similar to f in $[-4, 4]$.

Power Series



Example 5.4 (Taylor Series)
Consider the function in Figure 5.4 given by

$$f(x) = \sin(x) + \cos(x) \in \mathcal{C}^{\infty} . \tag{5.19}$$

We seek a Taylor series expansion of f at $x_0 = 0$, which is the Maclaurin series expansion of f. We obtain the following derivatives:

$$\begin{aligned}
f(0) &= \sin(0) + \cos(0) = 1 && (5.20)\\
f'(0) &= \cos(0) - \sin(0) = 1 && (5.21)\\
f''(0) &= -\sin(0) - \cos(0) = -1 && (5.22)\\
f^{(3)}(0) &= -\cos(0) + \sin(0) = -1 && (5.23)\\
f^{(4)}(0) &= \sin(0) + \cos(0) = f(0) = 1 && (5.24)\\
&\vdots
\end{aligned}$$

We can see a pattern here: The coefficients in our Taylor series are only ± 1 (since $\sin(0) = 0$), each of which occurs twice before switching to the other one. Furthermore, $f^{(k+4)}(0) = f^{(k)}(0)$.

Therefore, the full Taylor series expansion of f at $x_0 = 0$ is given by

$$\begin{aligned}
T_{\infty}(x) &= \sum_{k=0}^{\infty} \frac{f^{(k)}(x_0)}{k!}(x - x_0)^k && (5.25a)\\
&= 1 + x - \frac{1}{2!}x^2 - \frac{1}{3!}x^3 + \frac{1}{4!}x^4 + \frac{1}{5!}x^5 - \cdots && (5.25b)\\
&= 1 - \frac{1}{2!}x^2 + \frac{1}{4!}x^4 \mp \cdots + x - \frac{1}{3!}x^3 + \frac{1}{5!}x^5 \mp \cdots && (5.25c)\\
&= \sum_{k=0}^{\infty} (-1)^k \frac{1}{(2k)!} x^{2k} + \sum_{k=0}^{\infty} (-1)^k \frac{1}{(2k+1)!} x^{2k+1} && (5.25d)\\
&= \cos(x) + \sin(x) , && (5.25e)
\end{aligned}$$

where we used the *power series representations*

power series representation

$$\cos(x) = \sum_{k=0}^{\infty} (-1)^k \frac{1}{(2k)!} x^{2k}, \tag{5.26}$$

$$\sin(x) = \sum_{k=0}^{\infty} (-1)^k \frac{1}{(2k+1)!} x^{2k+1}. \tag{5.27}$$

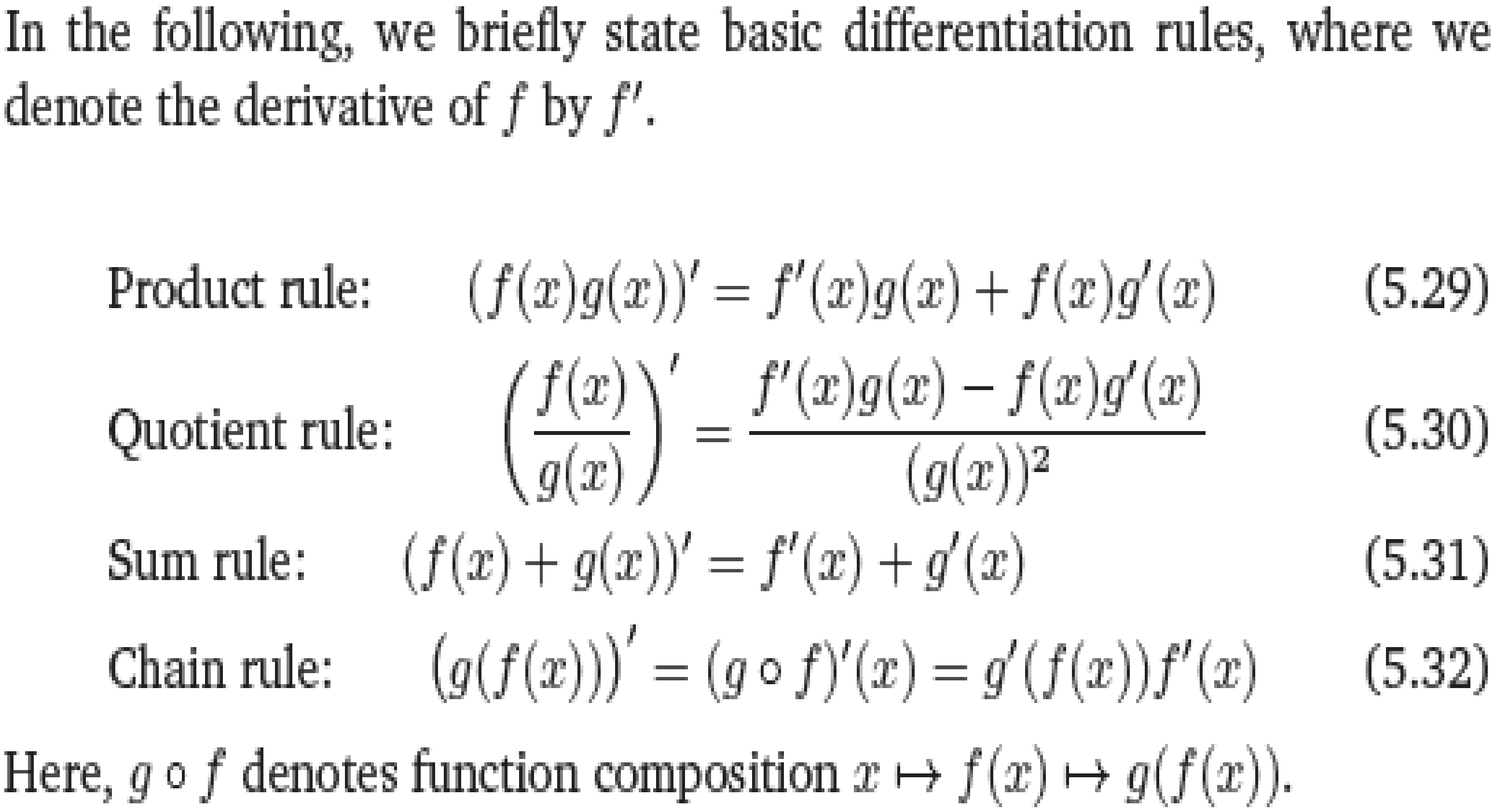
Figure 5.4 shows the corresponding first Taylor polynomials T_n for $n = 0, 1, 5, 10$.

Remark. A Taylor series is a special case of a power series

$$f(x) = \sum_{k=0}^{\infty} a_k (x - c)^k \tag{5.28}$$

where a_k are coefficients and c is a constant, which has the special form in Definition 5.4. ◇

Differentiation Rules



In the following, we briefly state basic differentiation rules, where we denote the derivative of f by f'.

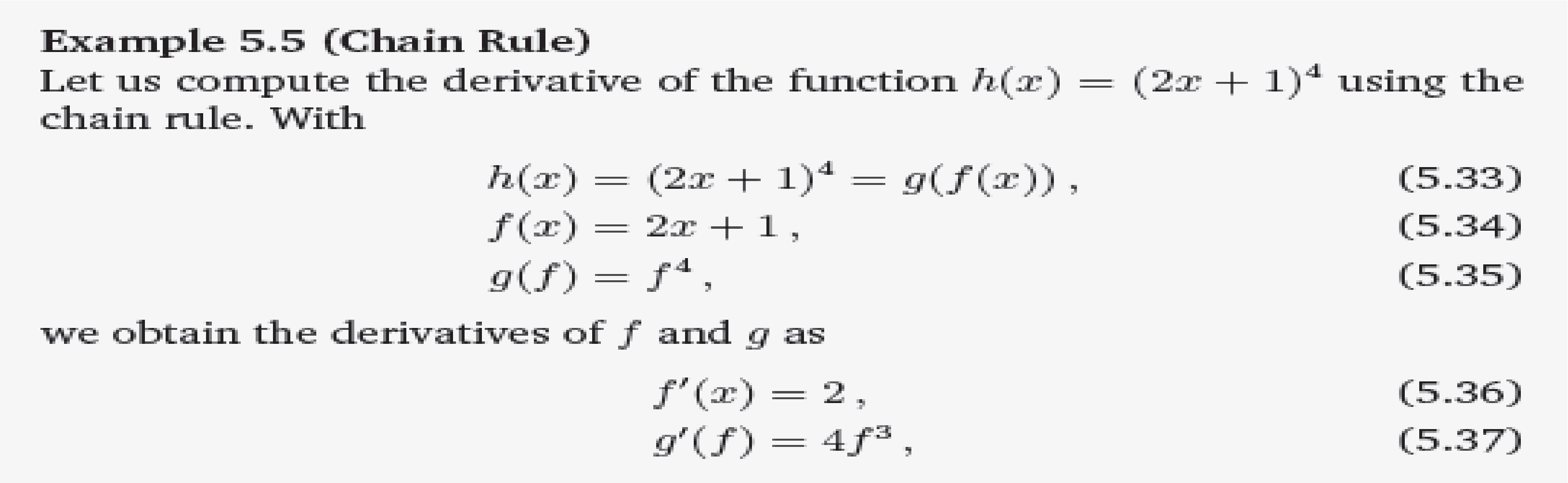
$$\text{Product rule:} \quad (f(x)g(x))' = f'(x)g(x) + f(x)g'(x) \tag{5.29}$$

$$\text{Quotient rule:} \quad \left(\frac{f(x)}{g(x)}\right)' = \frac{f'(x)g(x) - f(x)g'(x)}{(g(x))^2} \tag{5.30}$$

$$\text{Sum rule:} \quad (f(x) + g(x))' = f'(x) + g'(x) \tag{5.31}$$

$$\text{Chain rule:} \quad \big(g(f(x))\big)' = (g \circ f)'(x) = g'(f(x))f'(x) \tag{5.32}$$

Here, $g \circ f$ denotes function composition $x \mapsto f(x) \mapsto g(f(x))$.

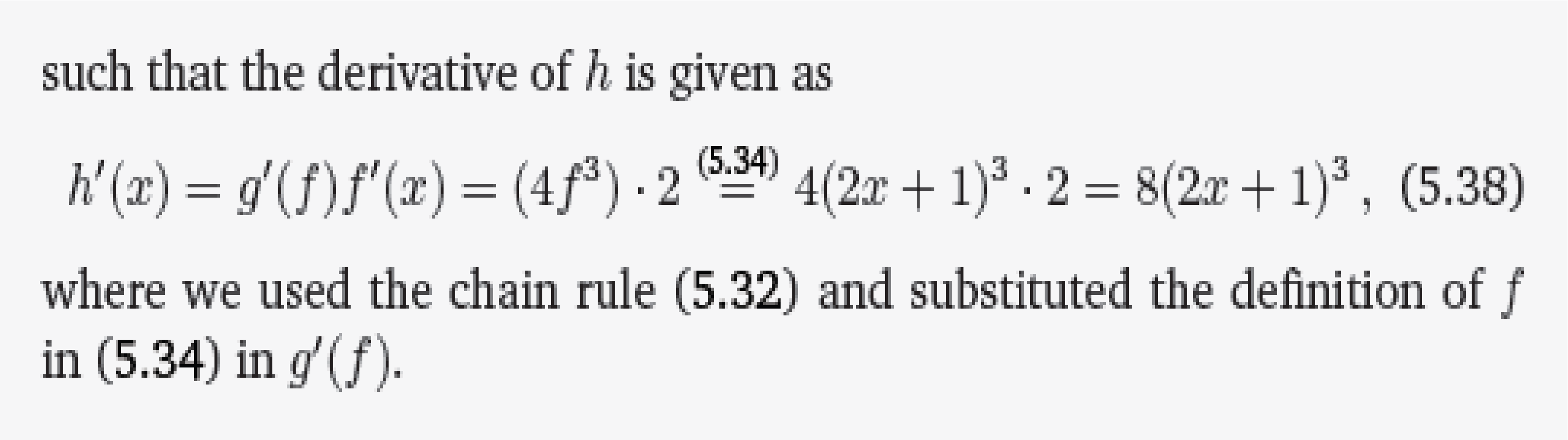


Example 5.5 (Chain Rule)
Let us compute the derivative of the function $h(x) = (2x+1)^4$ using the chain rule. With

$$h(x) = (2x+1)^4 = g(f(x))\,, \tag{5.33}$$
$$f(x) = 2x + 1\,, \tag{5.34}$$
$$g(f) = f^4\,, \tag{5.35}$$

we obtain the derivatives of f and g as

$$f'(x) = 2\,, \tag{5.36}$$
$$g'(f) = 4f^3\,, \tag{5.37}$$

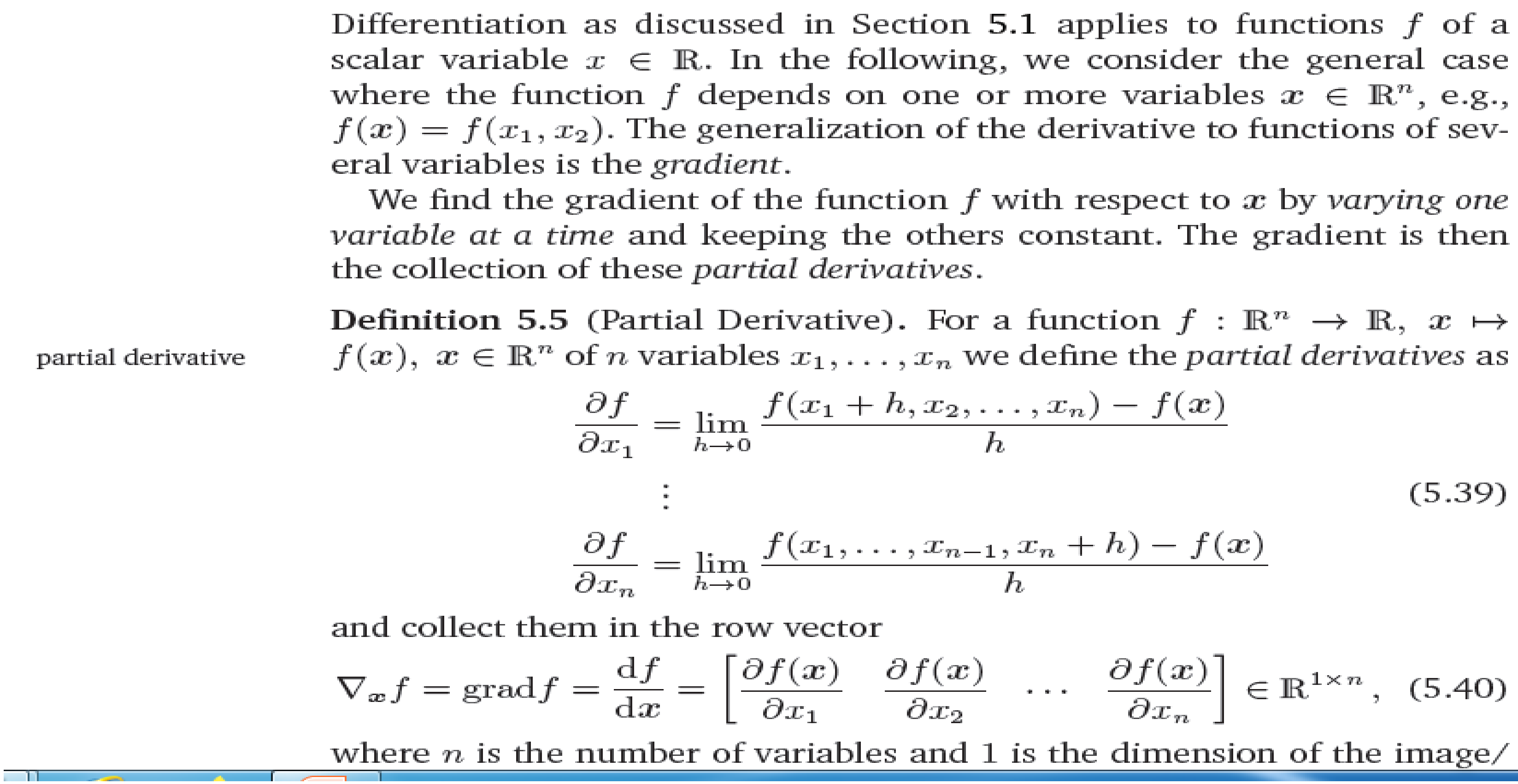


such that the derivative of h is given as

$$h'(x) = g'(f)f'(x) = (4f^3)\cdot 2 \overset{(5.34)}{=} 4(2x+1)^3 \cdot 2 = 8(2x+1)^3\,, \tag{5.38}$$

where we used the chain rule (5.32) and substituted the definition of f in (5.34) in $g'(f)$.

Partial Differentiation and Gradients



Differentiation as discussed in Section 5.1 applies to functions f of a scalar variable $x \in \mathbb{R}$. In the following, we consider the general case where the function f depends on one or more variables $\boldsymbol{x} \in \mathbb{R}^n$, e.g., $f(\boldsymbol{x}) = f(x_1, x_2)$. The generalization of the derivative to functions of several variables is the *gradient*.

We find the gradient of the function f with respect to $\boldsymbol{x}$ by *varying one variable at a time* and keeping the others constant. The gradient is then the collection of these *partial derivatives*.

partial derivative

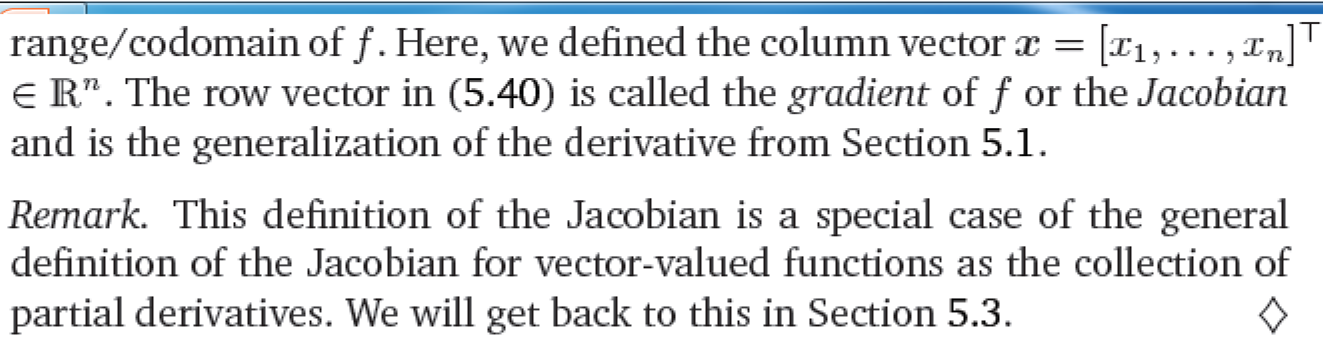
Definition 5.5 (Partial Derivative)**.** For a function $f : \mathbb{R}^n \to \mathbb{R}$, $\boldsymbol{x} \mapsto f(\boldsymbol{x})$, $\boldsymbol{x} \in \mathbb{R}^n$ of n variables $x_1, \dots, x_n$ we define the *partial derivatives* as

$$\begin{aligned} \frac{\partial f}{\partial x_1} &= \lim_{h \to 0} \frac{f(x_1 + h, x_2, \dots, x_n) - f(\boldsymbol{x})}{h} \\ &\vdots \\ \frac{\partial f}{\partial x_n} &= \lim_{h \to 0} \frac{f(x_1, \dots, x_{n-1}, x_n + h) - f(\boldsymbol{x})}{h} \end{aligned} \tag{5.39}$$

and collect them in the row vector

$$\nabla_{\boldsymbol{x}} f = \operatorname{grad} f = \frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}} = \begin{bmatrix} \frac{\partial f(\boldsymbol{x})}{\partial x_1} & \frac{\partial f(\boldsymbol{x})}{\partial x_2} & \cdots & \frac{\partial f(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \in \mathbb{R}^{1 \times n}, \tag{5.40}$$

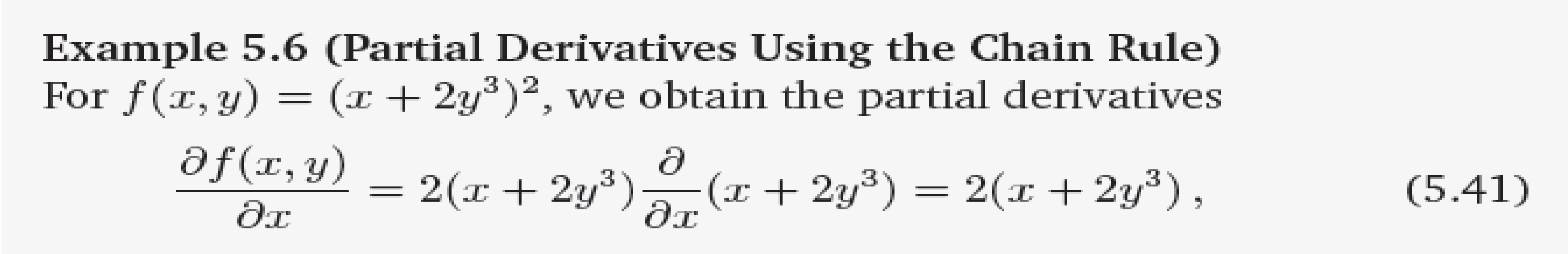
where n is the number of variables and 1 is the dimension of the image/



range/codomain of f. Here, we defined the column vector $\boldsymbol{x} = [x_1, \dots, x_n]^\top \in \mathbb{R}^n$. The row vector in (5.40) is called the *gradient* of f or the *Jacobian* and is the generalization of the derivative from Section 5.1.

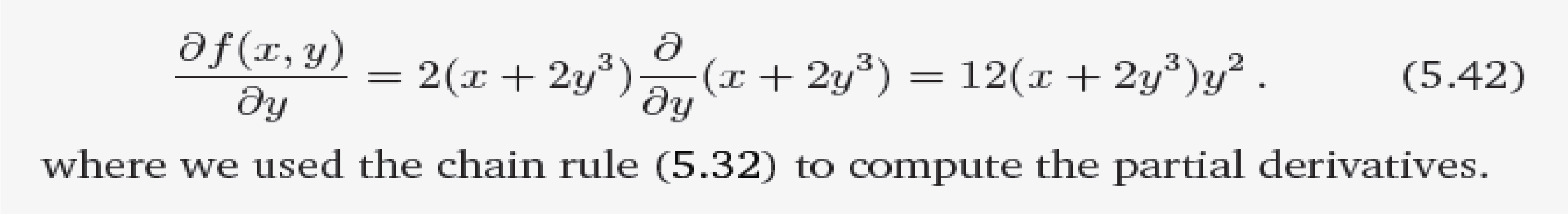
Remark. This definition of the Jacobian is a special case of the general definition of the Jacobian for vector-valued functions as the collection of partial derivatives. We will get back to this in Section 5.3. ◇

Partial Differentiation and Gradients



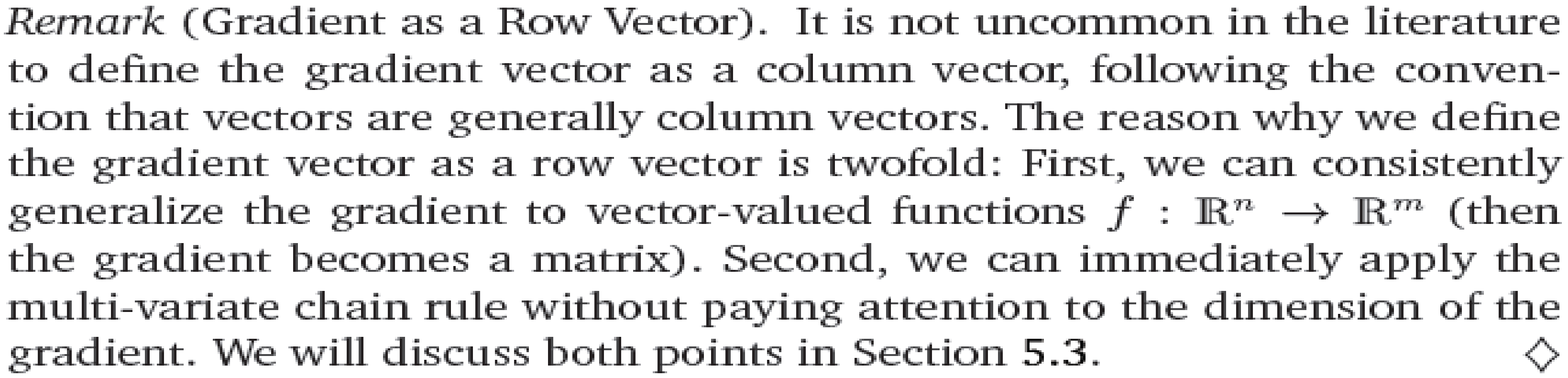
Example 5.6 (Partial Derivatives Using the Chain Rule)
For $f(x,y) = (x+2y^3)^2$, we obtain the partial derivatives

$$\frac{\partial f(x,y)}{\partial x} = 2(x+2y^3)\frac{\partial}{\partial x}(x+2y^3) = 2(x+2y^3)\,, \tag{5.41}$$



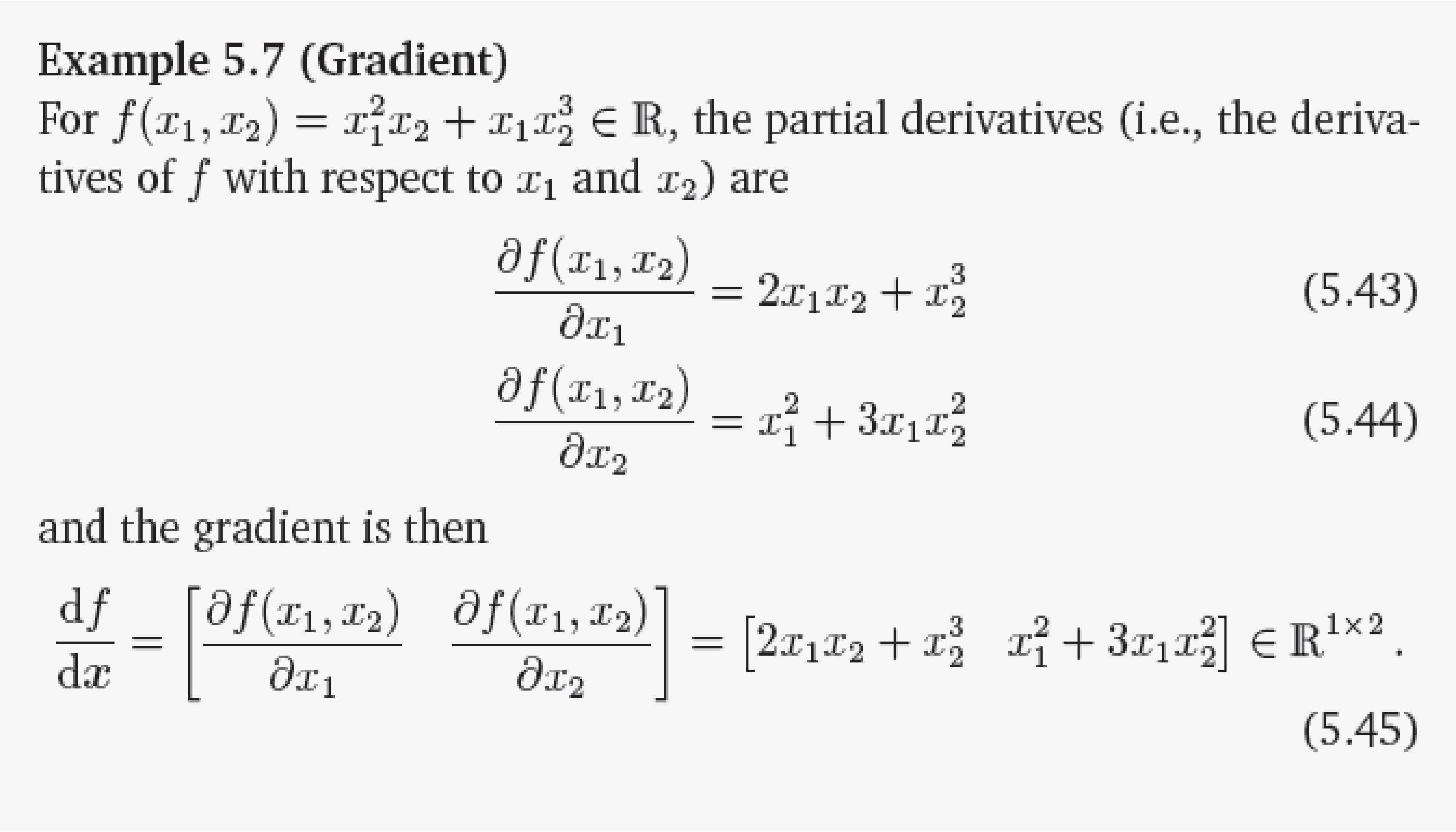
$$\frac{\partial f(x,y)}{\partial y} = 2(x+2y^3)\frac{\partial}{\partial y}(x+2y^3) = 12(x+2y^3)y^2\,. \tag{5.42}$$

where we used the chain rule (5.32) to compute the partial derivatives.



Remark (Gradient as a Row Vector). It is not uncommon in the literature to define the gradient vector as a column vector, following the convention that vectors are generally column vectors. The reason why we define the gradient vector as a row vector is twofold: First, we can consistently generalize the gradient to vector-valued functions $f : \mathbb{R}^n \to \mathbb{R}^m$ (then the gradient becomes a matrix). Second, we can immediately apply the multi-variate chain rule without paying attention to the dimension of the gradient. We will discuss both points in Section 5.3. ◇

Gradient



Example 5.7 (Gradient)

For $f(x_1, x_2) = x_1^2 x_2 + x_1 x_2^3 \in \mathbb{R}$, the partial derivatives (i.e., the derivatives of f with respect to x_1 and x_2) are

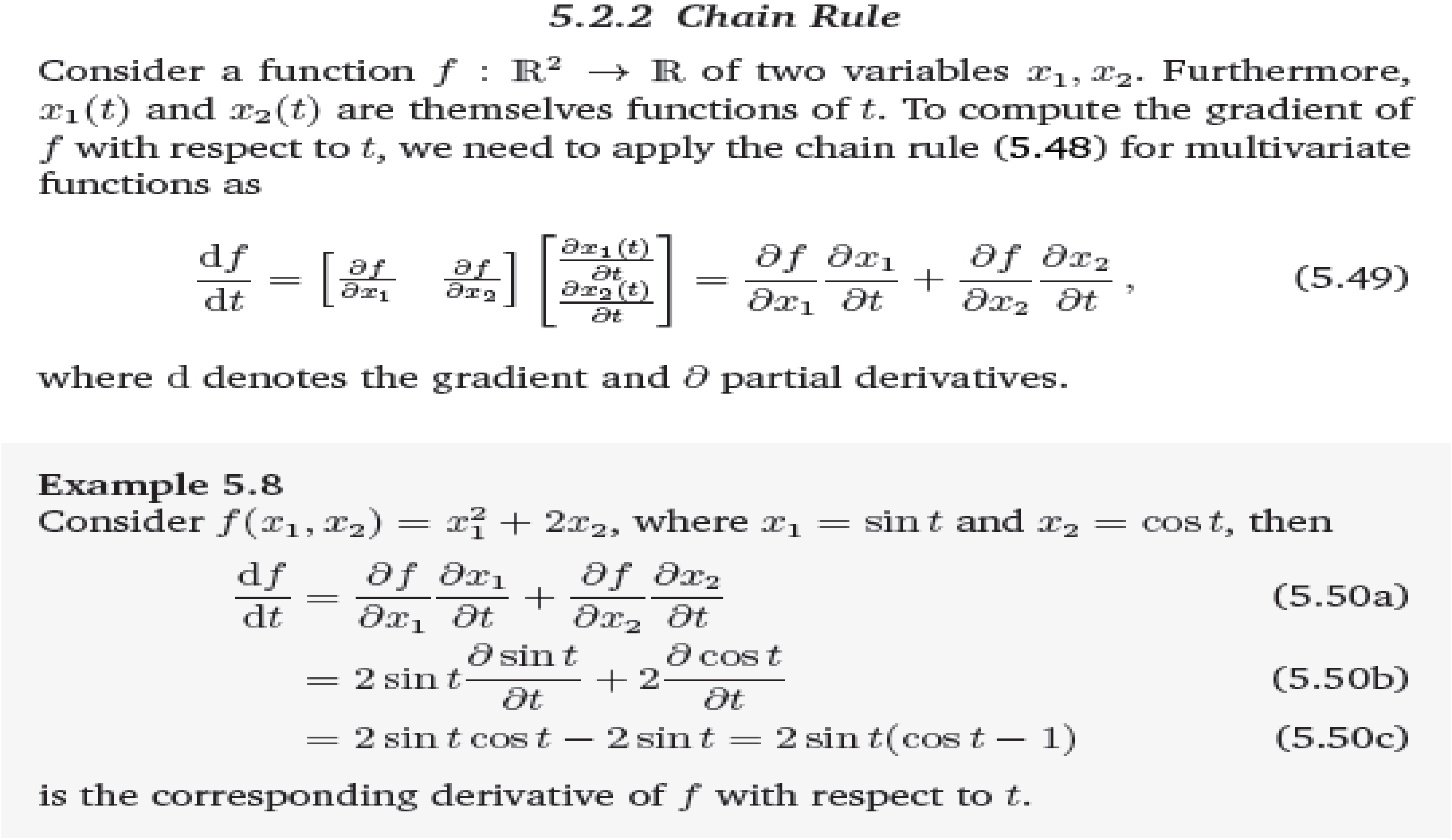
$$\frac{\partial f(x_1, x_2)}{\partial x_1} = 2x_1 x_2 + x_2^3 \tag{5.43}$$

$$\frac{\partial f(x_1, x_2)}{\partial x_2} = x_1^2 + 3x_1 x_2^2 \tag{5.44}$$

and the gradient is then

$$\frac{\mathrm{d}f}{\mathrm{d}\boldsymbol{x}} = \begin{bmatrix} \frac{\partial f(x_1, x_2)}{\partial x_1} & \frac{\partial f(x_1, x_2)}{\partial x_2} \end{bmatrix} = \begin{bmatrix} 2x_1 x_2 + x_2^3 & x_1^2 + 3x_1 x_2^2 \end{bmatrix} \in \mathbb{R}^{1\times 2} . \tag{5.45}$$

Chain Rules



5.2.2 *Chain Rule*

Consider a function $f : \mathbb{R}^2 \to \mathbb{R}$ of two variables x_1, x_2. Furthermore, $x_1(t)$ and $x_2(t)$ are themselves functions of t. To compute the gradient of f with respect to t, we need to apply the chain rule (5.48) for multivariate functions as

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix} \begin{bmatrix} \frac{\partial x_1(t)}{\partial t} \\ \frac{\partial x_2(t)}{\partial t} \end{bmatrix} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t}\,, \tag{5.49}$$

where d denotes the gradient and ∂ partial derivatives.

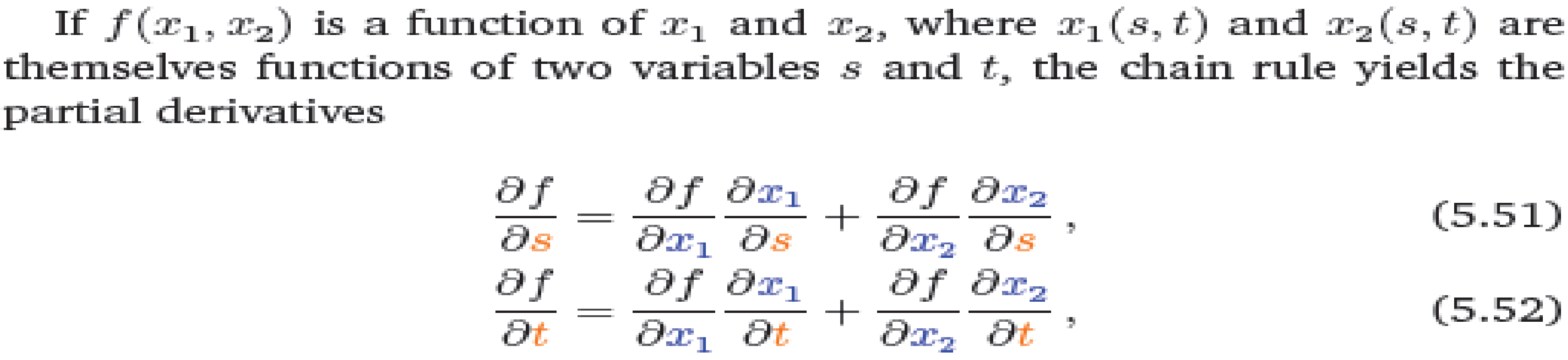
Example 5.8
Consider $f(x_1, x_2) = x_1^2 + 2x_2$, where $x_1 = \sin t$ and $x_2 = \cos t$, then

$$\frac{\mathrm{d}f}{\mathrm{d}t} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t} \tag{5.50a}$$

$$= 2\sin t\frac{\partial \sin t}{\partial t} + 2\frac{\partial \cos t}{\partial t} \tag{5.50b}$$

$$= 2\sin t\cos t - 2\sin t = 2\sin t(\cos t - 1) \tag{5.50c}$$

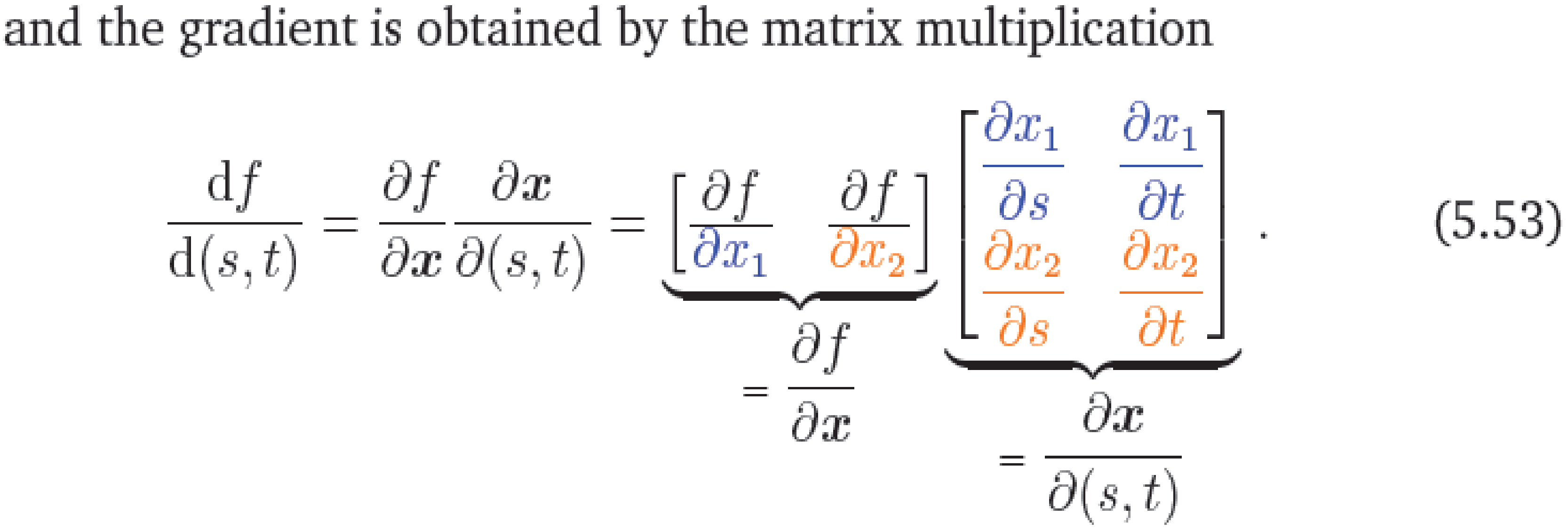
is the corresponding derivative of f with respect to t.



If $f(x_1, x_2)$ is a function of x_1 and x_2, where $x_1(s,t)$ and $x_2(s,t)$ are themselves functions of two variables s and t, the chain rule yields the partial derivatives

$$\frac{\partial f}{\partial s} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial s} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial s}\,, \tag{5.51}$$

$$\frac{\partial f}{\partial t} = \frac{\partial f}{\partial x_1}\frac{\partial x_1}{\partial t} + \frac{\partial f}{\partial x_2}\frac{\partial x_2}{\partial t}\,, \tag{5.52}$$



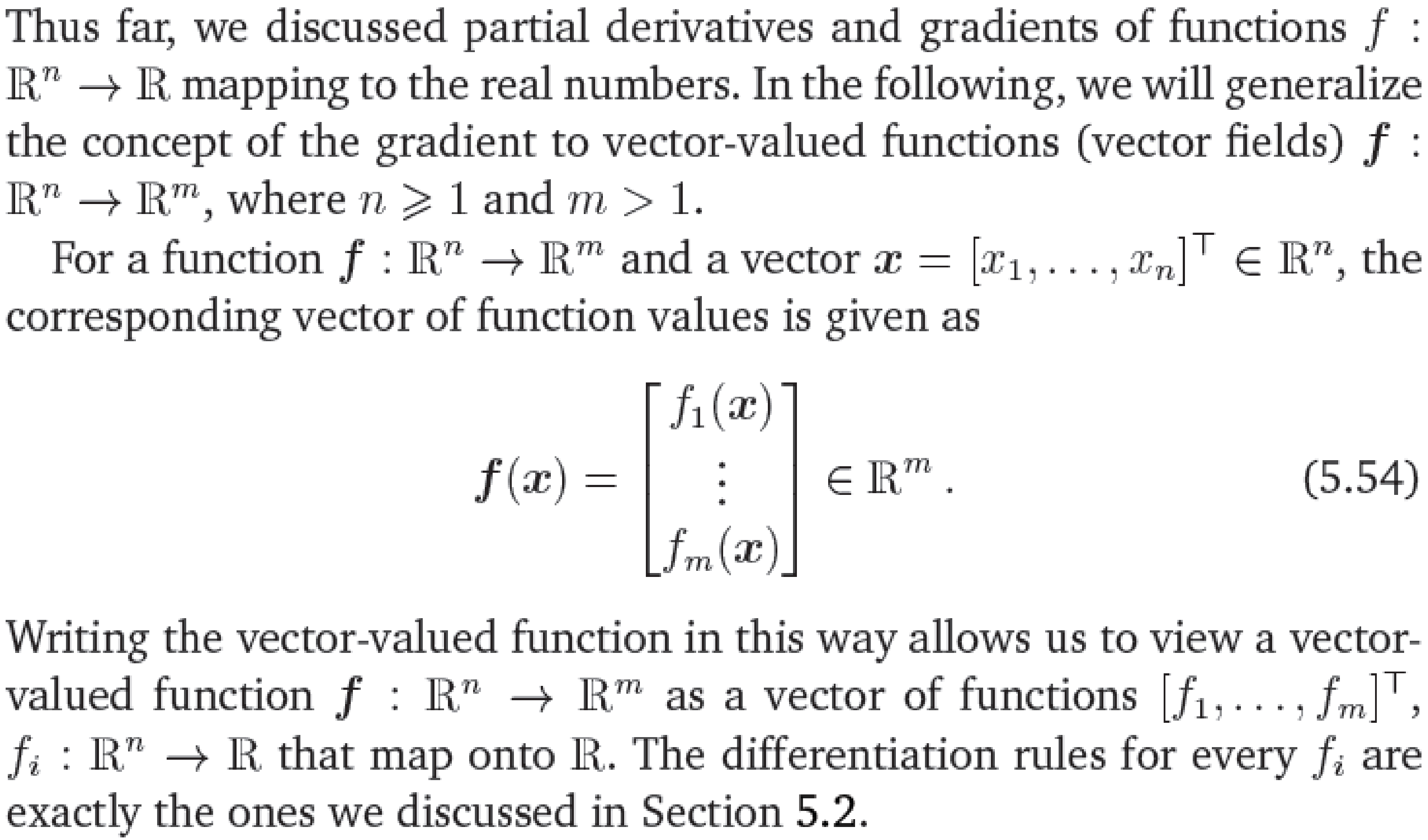
and the gradient is obtained by the matrix multiplication

$$\frac{\mathrm{d}f}{\mathrm{d}(s,t)} = \frac{\partial f}{\partial \boldsymbol{x}}\frac{\partial \boldsymbol{x}}{\partial(s,t)} = \underbrace{\begin{bmatrix} \frac{\partial f}{\partial x_1} & \frac{\partial f}{\partial x_2} \end{bmatrix}}_{=\frac{\partial f}{\partial \boldsymbol{x}}} \underbrace{\begin{bmatrix} \frac{\partial x_1}{\partial s} & \frac{\partial x_1}{\partial t} \\ \frac{\partial x_2}{\partial s} & \frac{\partial x_2}{\partial t} \end{bmatrix}}_{=\frac{\partial \boldsymbol{x}}{\partial(s,t)}} \,. \tag{5.53}$$

Remark (Verifying the Correctness of a Gradient Implementation). The definition of the partial derivatives as the limit of the corresponding difference quotient (see (5.39)) can be exploited when numerically checking the correctness of gradients in computer programs: When we compute gradients and implement them, we can use finite differences to numerically test our computation and implementation: We choose the value h to be small (e.g., $h = 10^{-4}$) and compare the finite-difference approximation from (5.39) with our (analytic) implementation of the gradient. If the error is small, our gradient implementation is probably correct. "Small" could mean that $\sqrt{\frac{\sum_i (dh_i - df_i)^2}{\sum_i (dh_i + df_i)^2}} < 10^{-6}$, where dh_i is the finite-difference approximation and df_i is the analytic gradient of f with respect to the ith variable x_i. $\diamondsuit$

Gradient checking

Gradients of Vector-Valued Functions



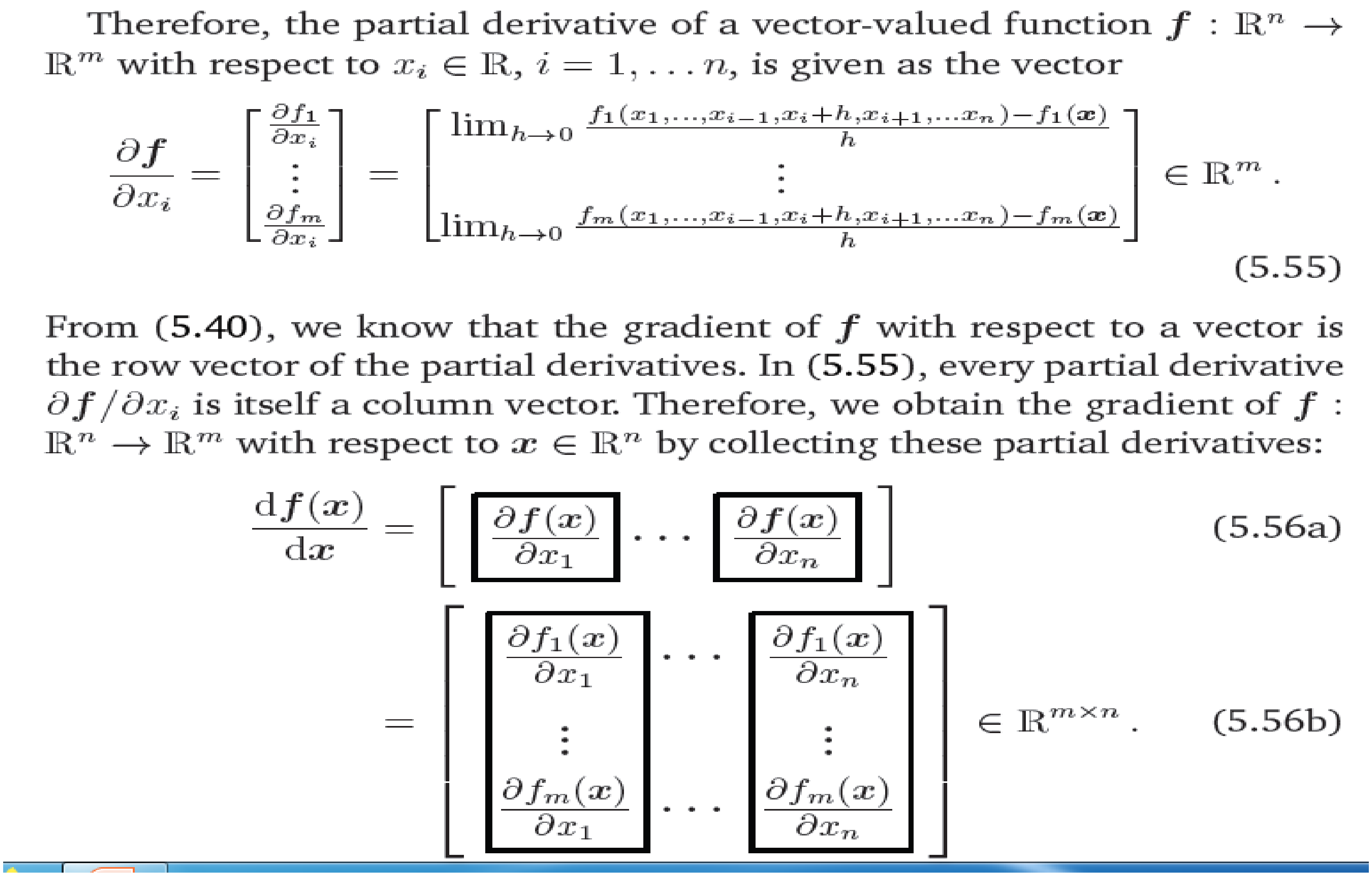
Thus far, we discussed partial derivatives and gradients of functions $f : \mathbb{R}^n \to \mathbb{R}$ mapping to the real numbers. In the following, we will generalize the concept of the gradient to vector-valued functions (vector fields) $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$, where $n \geqslant 1$ and $m > 1$.

For a function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ and a vector $\boldsymbol{x} = [x_1, \ldots, x_n]^\top \in \mathbb{R}^n$, the corresponding vector of function values is given as

$$\boldsymbol{f}(\boldsymbol{x}) = \begin{bmatrix} f_1(\boldsymbol{x}) \\ \vdots \\ f_m(\boldsymbol{x}) \end{bmatrix} \in \mathbb{R}^m \,. \tag{5.54}$$

Writing the vector-valued function in this way allows us to view a vector-valued function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ as a vector of functions $[f_1, \ldots, f_m]^\top$, $f_i : \mathbb{R}^n \to \mathbb{R}$ that map onto $\mathbb{R}$. The differentiation rules for every f_i are exactly the ones we discussed in Section 5.2.

Gradients of Vector-Valued Functions



Therefore, the partial derivative of a vector-valued function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ with respect to $x_i \in \mathbb{R}$, $i = 1, \dots n$, is given as the vector

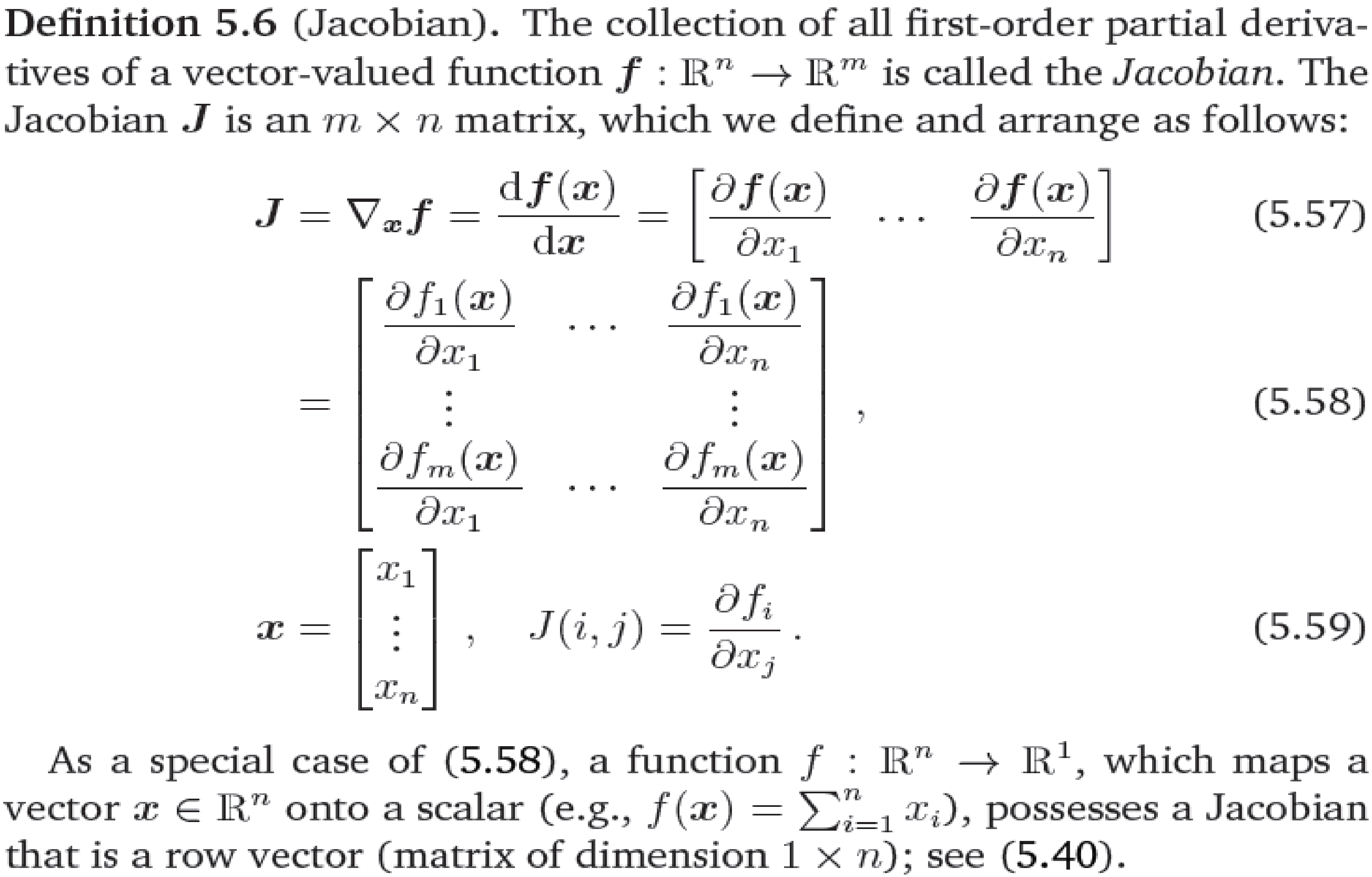
$$\frac{\partial \boldsymbol{f}}{\partial x_i} = \begin{bmatrix} \frac{\partial f_1}{\partial x_i} \\ \vdots \\ \frac{\partial f_m}{\partial x_i} \end{bmatrix} = \begin{bmatrix} \lim_{h \to 0} \frac{f_1(x_1, \dots, x_{i-1}, x_i + h, x_{i+1}, \dots x_n) - f_1(\boldsymbol{x})}{h} \\ \vdots \\ \lim_{h \to 0} \frac{f_m(x_1, \dots, x_{i-1}, x_i + h, x_{i+1}, \dots x_n) - f_m(\boldsymbol{x})}{h} \end{bmatrix} \in \mathbb{R}^m . \tag{5.55}$$

From (5.40), we know that the gradient of $\boldsymbol{f}$ with respect to a vector is the row vector of the partial derivatives. In (5.55), every partial derivative $\partial \boldsymbol{f} / \partial x_i$ is itself a column vector. Therefore, we obtain the gradient of $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ with respect to $\boldsymbol{x} \in \mathbb{R}^n$ by collecting these partial derivatives:

$$\frac{\mathrm{d}\boldsymbol{f}(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}} = \begin{bmatrix} \boxed{\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_1}} & \cdots & \boxed{\frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_n}} \end{bmatrix} \tag{5.56a}$$

$$= \begin{bmatrix} \boxed{\begin{matrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} \\ \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} \end{matrix}} & \begin{matrix} \cdots \\ \\ \cdots \end{matrix} & \boxed{\begin{matrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{matrix}} \end{bmatrix} \in \mathbb{R}^{m \times n} . \tag{5.56b}$$

Jacobian



Definition 5.6 (Jacobian). The collection of all first-order partial derivatives of a vector-valued function $\boldsymbol{f} : \mathbb{R}^n \to \mathbb{R}^m$ is called the *Jacobian*. The Jacobian $\boldsymbol{J}$ is an $m \times n$ matrix, which we define and arrange as follows:

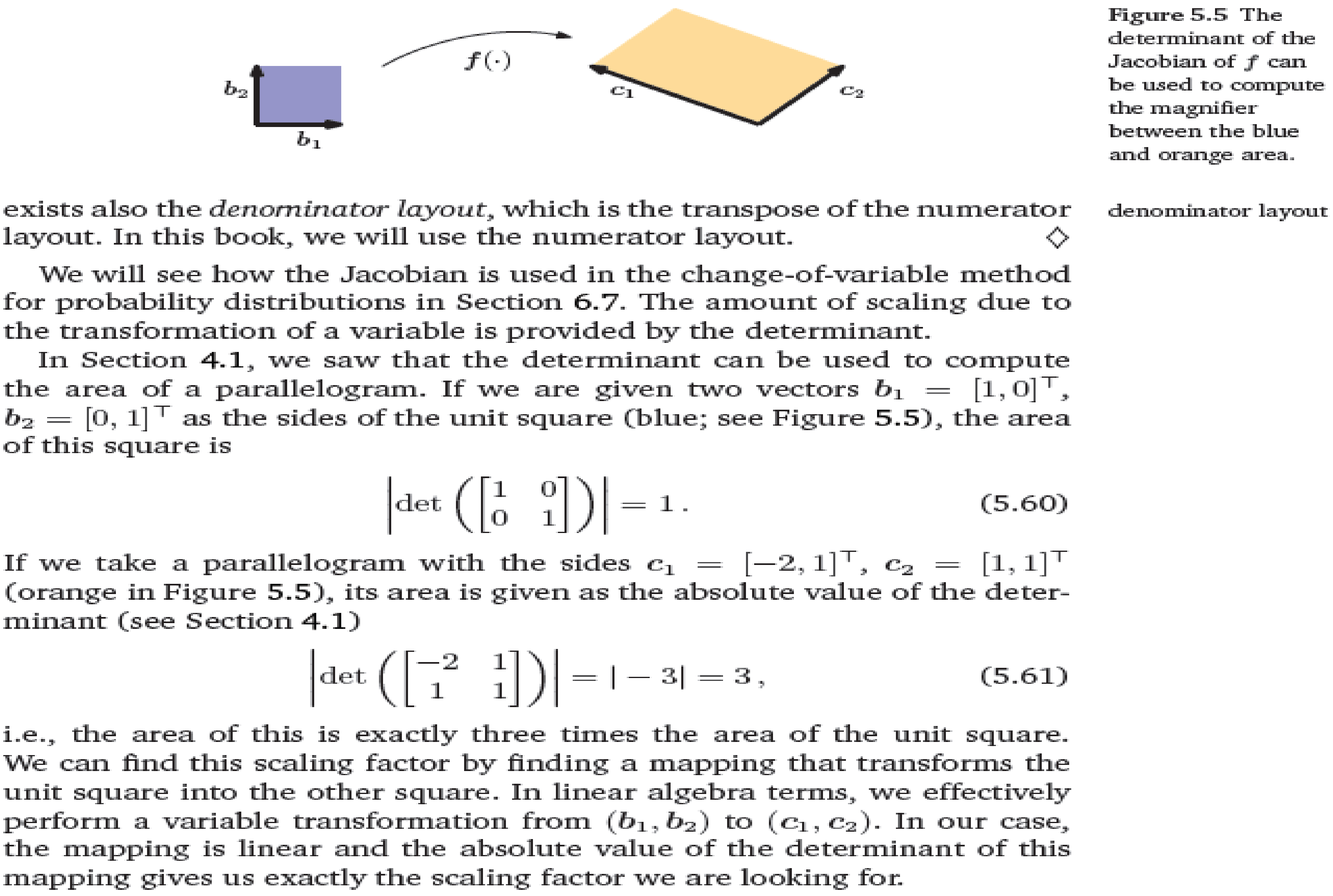
$$\boldsymbol{J} = \nabla_{\boldsymbol{x}}\boldsymbol{f} = \frac{\mathrm{d}\boldsymbol{f}(\boldsymbol{x})}{\mathrm{d}\boldsymbol{x}} = \begin{bmatrix} \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial \boldsymbol{f}(\boldsymbol{x})}{\partial x_n} \end{bmatrix} \tag{5.57}$$

$$= \begin{bmatrix} \frac{\partial f_1(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_1(\boldsymbol{x})}{\partial x_n} \\ \vdots & & \vdots \\ \frac{\partial f_m(\boldsymbol{x})}{\partial x_1} & \cdots & \frac{\partial f_m(\boldsymbol{x})}{\partial x_n} \end{bmatrix}, \tag{5.58}$$

$$\boldsymbol{x} = \begin{bmatrix} x_1 \\ \vdots \\ x_n \end{bmatrix}, \quad J(i,j) = \frac{\partial f_i}{\partial x_j}. \tag{5.59}$$

As a special case of (5.58), a function $f : \mathbb{R}^n \to \mathbb{R}^1$, which maps a vector $\boldsymbol{x} \in \mathbb{R}^n$ onto a scalar (e.g., $f(\boldsymbol{x}) = \sum_{i=1}^n x_i$), possesses a Jacobian that is a row vector (matrix of dimension $1 \times n$); see (5.40).

Jacobian is used to compute area of parallelogram



b_2
b_1
$f(\cdot)$
c_1
c_2

Figure 5.5 The determinant of the Jacobian of f can be used to compute the magnifier between the blue and orange area.

denominator layout

exists also the *denominator layout*, which is the transpose of the numerator layout. In this book, we will use the numerator layout. ◇

We will see how the Jacobian is used in the change-of-variable method for probability distributions in Section 6.7. The amount of scaling due to the transformation of a variable is provided by the determinant.

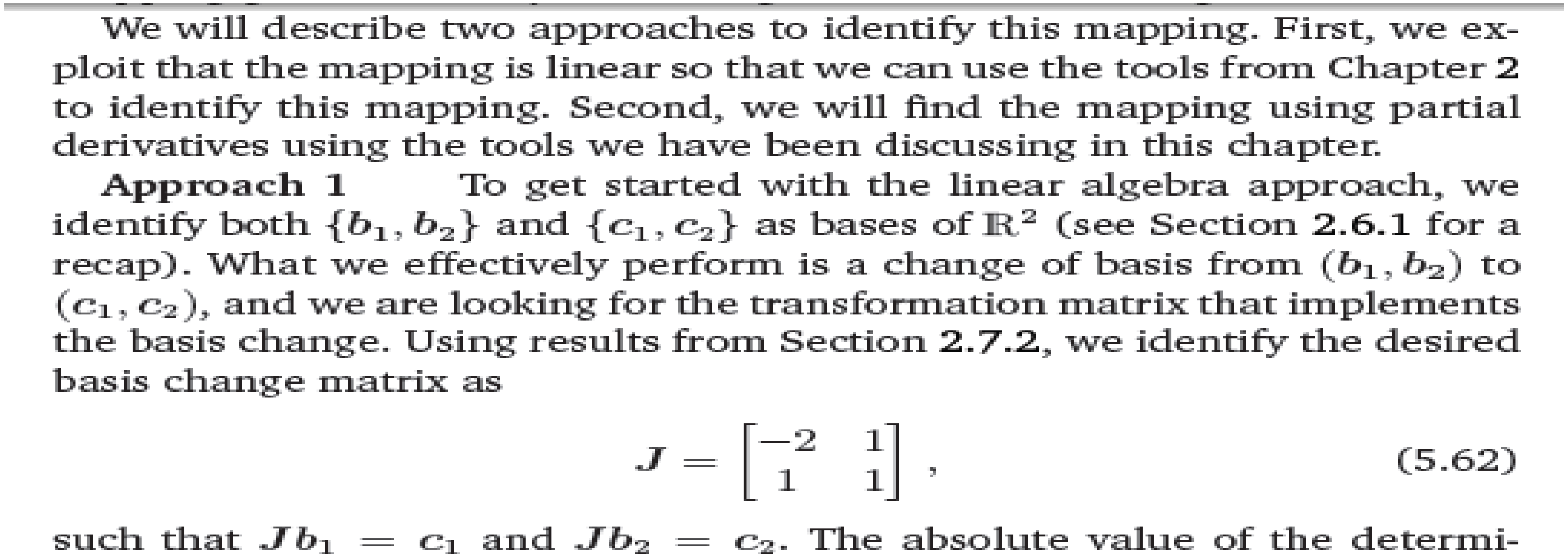
In Section 4.1, we saw that the determinant can be used to compute the area of a parallelogram. If we are given two vectors $b_1 = [1, 0]^\top$, $b_2 = [0, 1]^\top$ as the sides of the unit square (blue; see Figure 5.5), the area of this square is

$$\left|\det\left(\begin{bmatrix}1 & 0\\ 0 & 1\end{bmatrix}\right)\right| = 1\,. \tag{5.60}$$

If we take a parallelogram with the sides $c_1 = [-2, 1]^\top$, $c_2 = [1, 1]^\top$ (orange in Figure 5.5), its area is given as the absolute value of the determinant (see Section 4.1)

$$\left|\det\left(\begin{bmatrix}-2 & 1\\ 1 & 1\end{bmatrix}\right)\right| = |-3| = 3\,, \tag{5.61}$$

i.e., the area of this is exactly three times the area of the unit square. We can find this scaling factor by finding a mapping that transforms the unit square into the other square. In linear algebra terms, we effectively perform a variable transformation from (b_1, b_2) to (c_1, c_2). In our case, the mapping is linear and the absolute value of the determinant of this mapping gives us exactly the scaling factor we are looking for.



We will describe two approaches to identify this mapping. First, we exploit that the mapping is linear so that we can use the tools from Chapter 2 to identify this mapping. Second, we will find the mapping using partial derivatives using the tools we have been discussing in this chapter.

Approach 1 To get started with the linear algebra approach, we identify both $\{\boldsymbol{b}_1, \boldsymbol{b}_2\}$ and $\{\boldsymbol{c}_1, \boldsymbol{c}_2\}$ as bases of $\mathbb{R}^2$ (see Section 2.6.1 for a recap). What we effectively perform is a change of basis from $(\boldsymbol{b}_1, \boldsymbol{b}_2)$ to $(\boldsymbol{c}_1, \boldsymbol{c}_2)$, and we are looking for the transformation matrix that implements the basis change. Using results from Section 2.7.2, we identify the desired basis change matrix as

$$\boldsymbol{J} = \begin{bmatrix} -2 & 1 \\ 1 & 1 \end{bmatrix}, \tag{5.62}$$

such that $\boldsymbol{J}\boldsymbol{b}_1 = \boldsymbol{c}_1$ and $\boldsymbol{J}\boldsymbol{b}_2 = \boldsymbol{c}_2$. The absolute value of the determinant of $\boldsymbol{J}$, which yields the scaling factor we are looking for, is given as $|\det(\boldsymbol{J})| = 3$, i.e., the area of the square spanned by $(\boldsymbol{c}_1, \boldsymbol{c}_2)$ is three times greater than the area spanned by $(\boldsymbol{b}_1, \boldsymbol{b}_2)$.

Approach 2 The linear algebra approach works for linear transformations; for nonlinear transformations (which become relevant in Section 6.7), we follow a more general approach using partial derivatives.

For this approach, we consider a function $\boldsymbol{f} : \mathbb{R}^2 \to \mathbb{R}^2$ that performs a variable transformation. In our example, $\boldsymbol{f}$ maps the coordinate representation of any vector $\boldsymbol{x} \in \mathbb{R}^2$ with respect to $(\boldsymbol{b}_1, \boldsymbol{b}_2)$ onto the coordinate representation $\boldsymbol{y} \in \mathbb{R}^2$ with respect to $(\boldsymbol{c}_1, \boldsymbol{c}_2)$. We want to identify the mapping so that we can compute how an area (or volume) changes when it is being transformed by $\boldsymbol{f}$. For this, we need to find out how $\boldsymbol{f}(\boldsymbol{x})$ changes if we modify $\boldsymbol{x}$ a bit. This question is exactly answered by the Jacobian matrix $\frac{\mathrm{d}\boldsymbol{f}}{\mathrm{d}\boldsymbol{x}} \in \mathbb{R}^{2\times 2}$. Since we can write

$$y_1 = -2x_1 + x_2 \tag{5.63}$$

$$y_2 = x_1 + x_2 \tag{5.64}$$

we obtain the functional relationship between $\boldsymbol{x}$ and $\boldsymbol{y}$, which allows us to get the partial derivatives

$$\frac{\partial y_1}{\partial x_1} = -2\,, \quad \frac{\partial y_1}{\partial x_2} = 1\,, \quad \frac{\partial y_2}{\partial x_1} = 1\,, \quad \frac{\partial y_2}{\partial x_2} = 1 \tag{5.65}$$

and compose the Jacobian as

$$\boldsymbol{J} = \begin{bmatrix} \frac{\partial y_1}{\partial x_1} & \frac{\partial y_1}{\partial x_2} \\ \frac{\partial y_2}{\partial x_1} & \frac{\partial y_2}{\partial x_2} \end{bmatrix} = \begin{bmatrix} -2 & 1 \\ 1 & 1 \end{bmatrix} . \tag{5.66}$$

Geometrically, the Jacobian determinant gives the magnification/ scaling factor when we transform an area or volume.

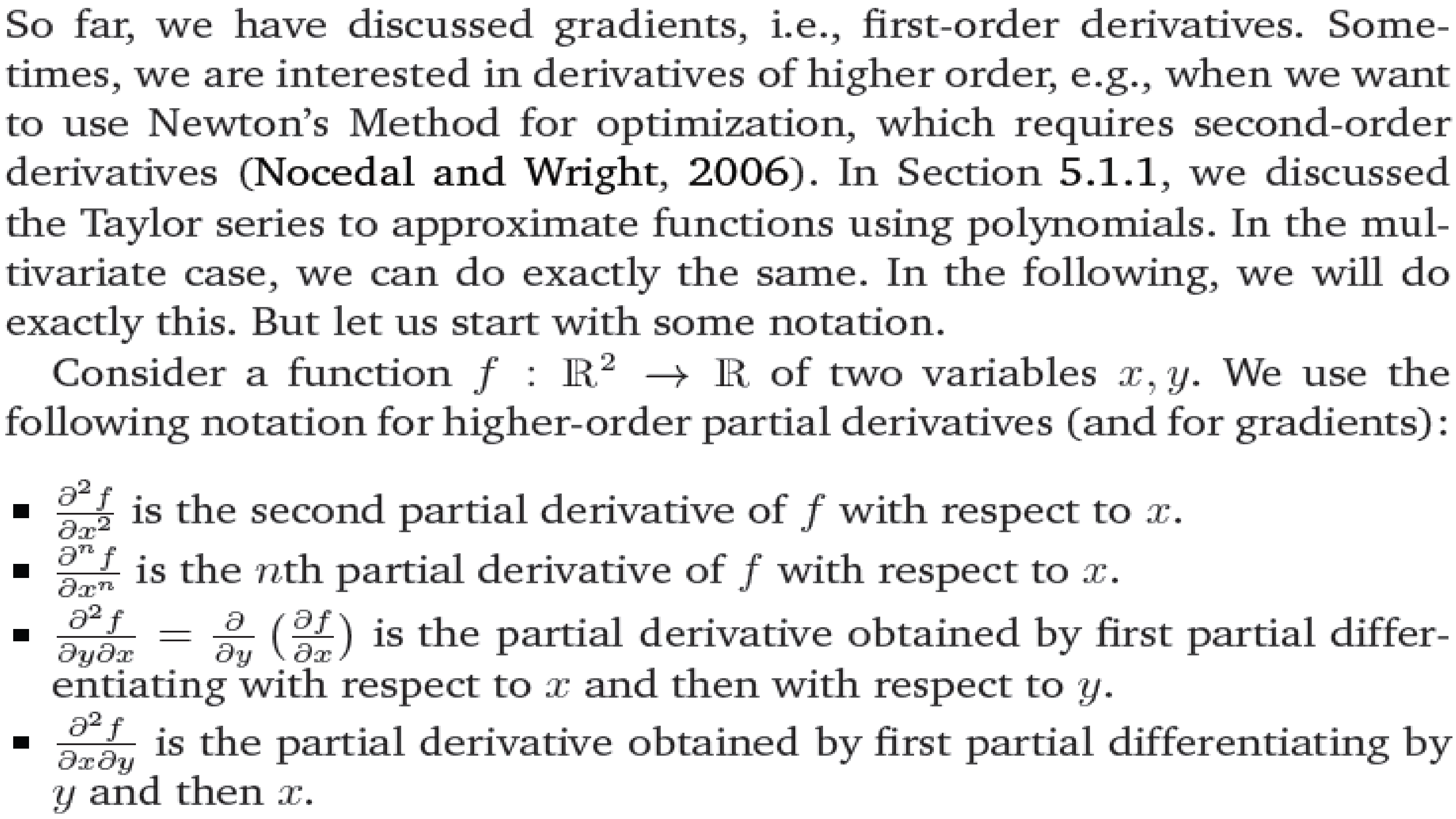
Jacobian determinant

The Jacobian represents the coordinate transformation we are looking for. It is exact if the coordinate transformation is linear (as in our case), and (5.66) recovers exactly the basis change matrix in (5.62). If the coordinate transformation is nonlinear, the Jacobian approximates this nonlinear transformation locally with a linear one. The absolute value of the *Jacobian determinant* $|\mathrm{det}(\boldsymbol{J})|$ is the factor by which areas or volumes are scaled when coordinates are transformed. Our case yields $|\mathrm{det}(\boldsymbol{J})| = 3$.

The Jacobian determinant and variable transformations will become relevant in Section 6.7 when we transform random variables and probability distributions. These transformations are extremely relevant in machine learning in the context of training deep neural networks using the *reparametrization trick*, also called *infinite perturbation analysis*.

Figure 5.6 Dimensionality of (partial) derivatives.

Higher-Order Derivative



So far, we have discussed gradients, i.e., first-order derivatives. Sometimes, we are interested in derivatives of higher order, e.g., when we want to use Newton's Method for optimization, which requires second-order derivatives (Nocedal and Wright, 2006). In Section 5.1.1, we discussed the Taylor series to approximate functions using polynomials. In the multivariate case, we can do exactly the same. In the following, we will do exactly this. But let us start with some notation.

Consider a function $f : \mathbb{R}^2 \to \mathbb{R}$ of two variables x, y. We use the following notation for higher-order partial derivatives (and for gradients):

- $\frac{\partial^2 f}{\partial x^2}$ is the second partial derivative of f with respect to x.
- $\frac{\partial^n f}{\partial x^n}$ is the nth partial derivative of f with respect to x.
- $\frac{\partial^2 f}{\partial y \partial x} = \frac{\partial}{\partial y}\left(\frac{\partial f}{\partial x}\right)$ is the partial derivative obtained by first partial differentiating with respect to x and then with respect to y.
- $\frac{\partial^2 f}{\partial x \partial y}$ is the partial derivative obtained by first partial differentiating by y and then x.

The Hessian

The *Hessian* is the collection of all second-order partial derivatives.

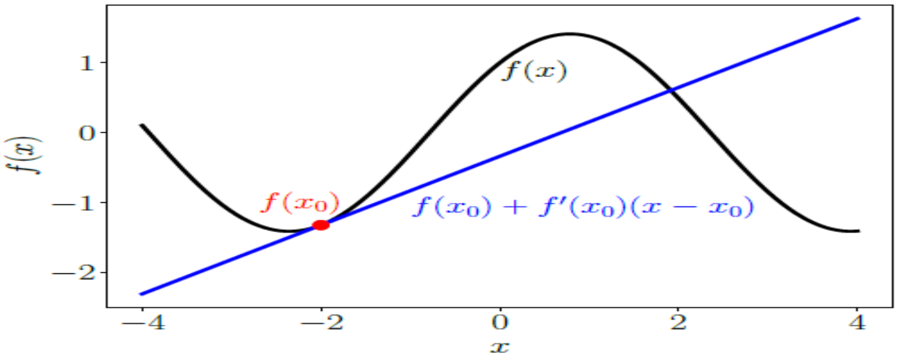


Figure 5.12 Linear approximation of a function. The original function f is linearized at $x_0 = -2$ using a first-order Taylor series expansion.

If $f(x, y)$ is a twice (continuously) differentiable function, then

$$\frac{\partial^2 f}{\partial x \partial y} = \frac{\partial^2 f}{\partial y \partial x}, \tag{5.146}$$

i.e., the order of differentiation does not matter, and the corresponding *Hessian matrix*

Hessian matrix

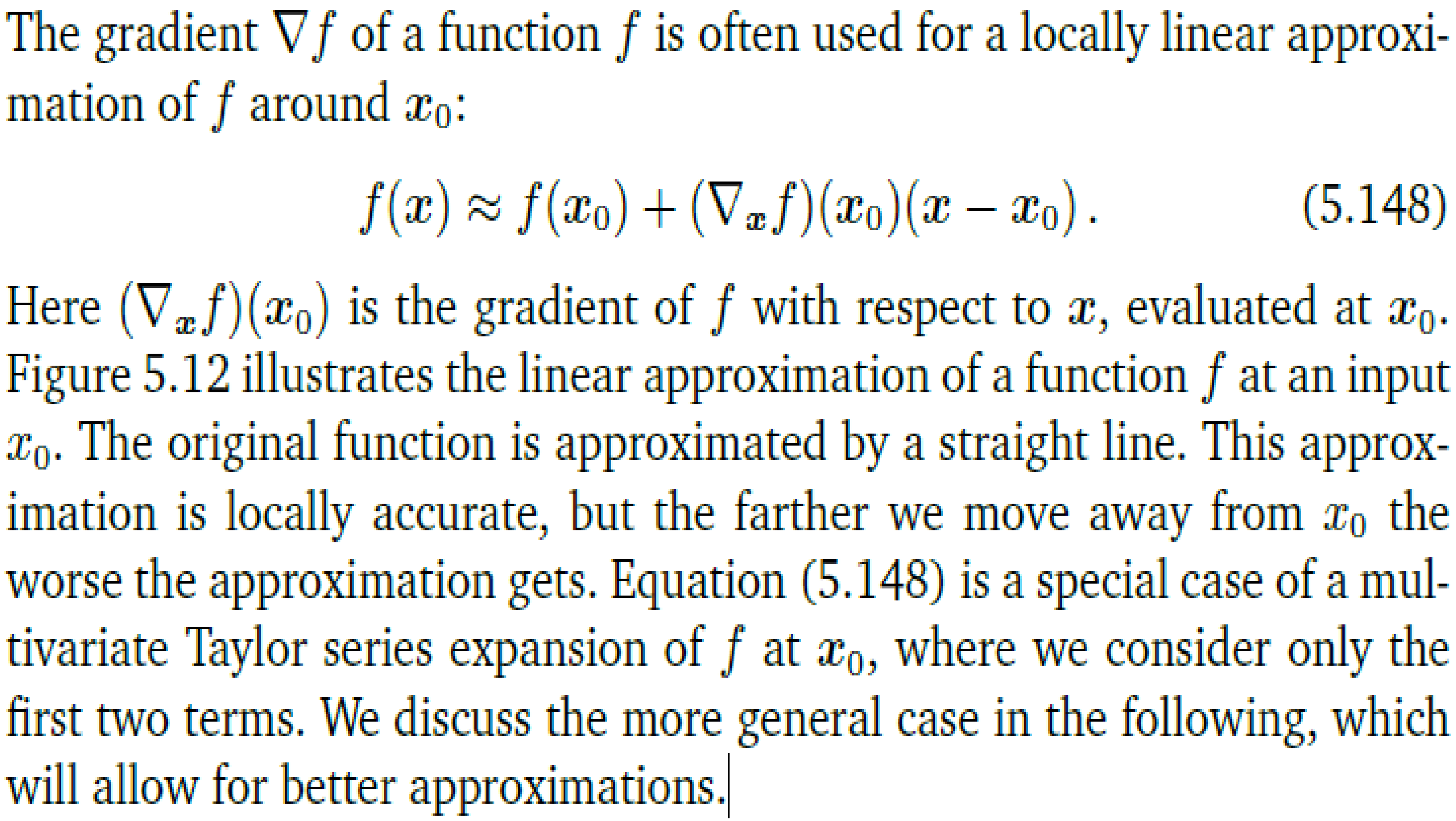
Hessian matrix

$$\boldsymbol{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial x \partial y} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} \tag{5.147}$$

is symmetric. The Hessian is denoted as $\nabla^2_{x,y} f(x, y)$. Generally, for $\boldsymbol{x} \in \mathbb{R}^n$ and $f : \mathbb{R}^n \to \mathbb{R}$, the Hessian is an $n \times n$ matrix. The Hessian measures the curvature of the function locally around (x, y).

Remark (Hessian of a Vector Field). If $f : \mathbb{R}^n \to \mathbb{R}^m$ is a vector field, the Hessian is an $(m \times n \times n)$-tensor. ◇

Linearization and Multivariate Taylor Series

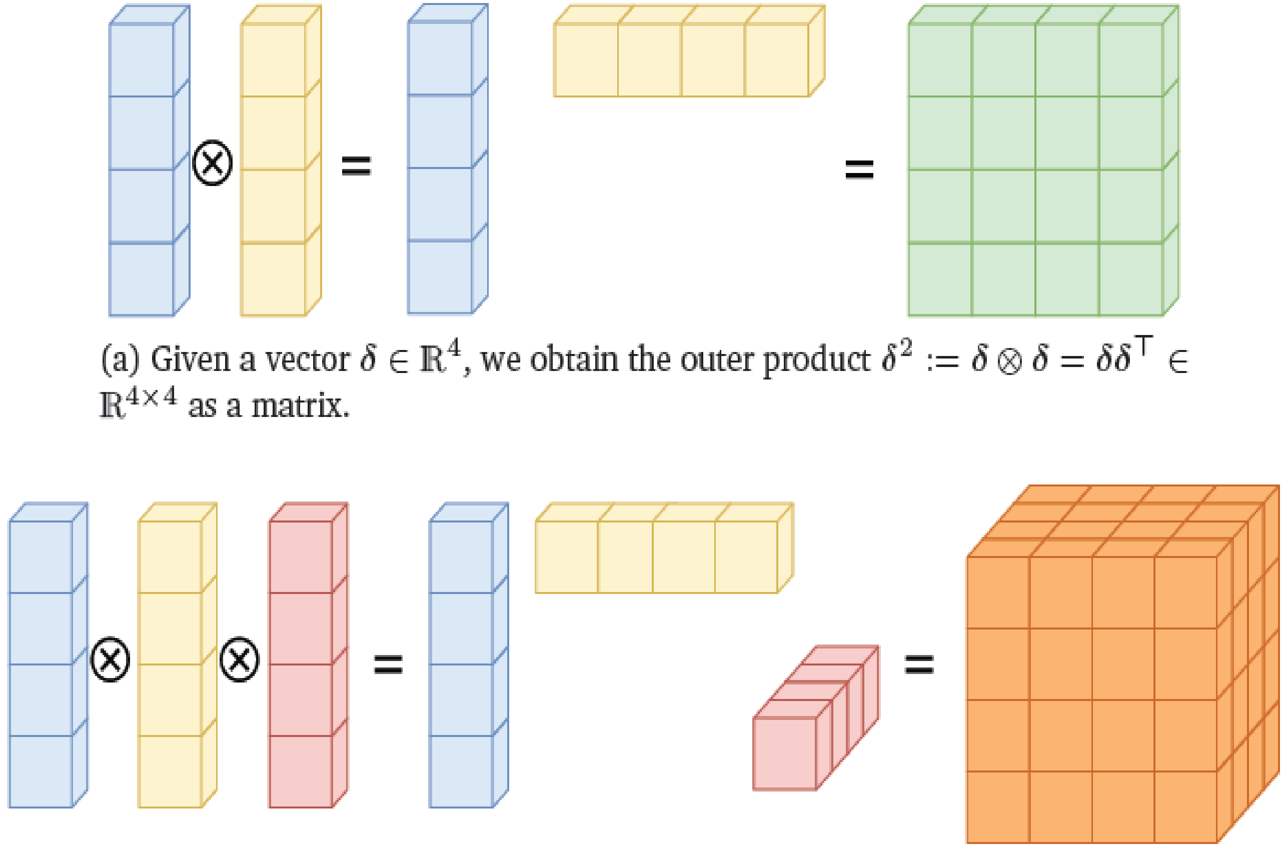


The gradient ∇f of a function f is often used for a locally linear approximation of f around $\boldsymbol{x}_0$:

$$f(\boldsymbol{x}) \approx f(\boldsymbol{x}_0) + (\nabla_{\boldsymbol{x}} f)(\boldsymbol{x}_0)(\boldsymbol{x} - \boldsymbol{x}_0)\,. \tag{5.148}$$

Here $(\nabla_{\boldsymbol{x}} f)(\boldsymbol{x}_0)$ is the gradient of f with respect to $\boldsymbol{x}$, evaluated at $\boldsymbol{x}_0$. Figure 5.12 illustrates the linear approximation of a function f at an input x_0. The original function is approximated by a straight line. This approximation is locally accurate, but the farther we move away from x_0 the worse the approximation gets. Equation (5.148) is a special case of a multivariate Taylor series expansion of f at $\boldsymbol{x}_0$, where we consider only the first two terms. We discuss the more general case in the following, which will allow for better approximations.

Figure 5.13 Visualizing outer products. Outer products of vectors increase the dimensionality of the array by 1 per term. (a) The outer product of two vectors results in a matrix; (b) the outer product of three vectors yields a third-order tensor.



(a) Given a vector $\delta \in \mathbb{R}^4$, we obtain the outer product $\delta^2 := \delta \otimes \delta = \delta\delta^\top \in \mathbb{R}^{4\times 4}$ as a matrix.

(b) An outer product $\delta^3 := \delta \otimes \delta \otimes \delta \in \mathbb{R}^{4\times 4\times 4}$ results in a third-order tensor ("three-dimensional matrix"), i.e., an array with three indexes.

Definition 5.7 (Multivariate Taylor Series)**.** We consider a function

$$f : \mathbb{R}^D \to \mathbb{R} \tag{5.149}$$

$$\boldsymbol{x} \mapsto f(\boldsymbol{x}) , \quad \boldsymbol{x} \in \mathbb{R}^D , \tag{5.150}$$

multivariate Taylor series

that is smooth at $\boldsymbol{x}_0$. When we define the difference vector $\boldsymbol{\delta} := \boldsymbol{x} - \boldsymbol{x}_0$, the *multivariate Taylor series* of f at $(\boldsymbol{x}_0)$ is defined as

$$f(\boldsymbol{x}) = \sum_{k=0}^{\infty} \frac{D_{\boldsymbol{x}}^k f(\boldsymbol{x}_0)}{k!} \boldsymbol{\delta}^k , \tag{5.151}$$

where $D_{\boldsymbol{x}}^k f(\boldsymbol{x}_0)$ is the k-th (total) derivative of f with respect to $\boldsymbol{x}$, evaluated at $\boldsymbol{x}_0$.

Taylor polynomial

Definition 5.8 (Taylor Polynomial)**.** The *Taylor polynomial* of degree n of f at $\boldsymbol{x}_0$ contains the first $n+1$ components of the series in (5.151) and is defined as

$$T_n(\boldsymbol{x}) = \sum_{k=0}^{n} \frac{D_{\boldsymbol{x}}^k f(\boldsymbol{x}_0)}{k!} \boldsymbol{\delta}^k . \tag{5.152}$$

In (5.151) and (5.152), we used the slightly sloppy notation of $\boldsymbol{\delta}^k$, which is not defined for vectors $\boldsymbol{x} \in \mathbb{R}^D$, $D > 1$, and $k > 1$. Note that both $D_{\boldsymbol{x}}^k f$ and $\boldsymbol{\delta}^k$ are k-th order tensors, i.e., k-dimensional arrays. The kth-order tensor $\boldsymbol{\delta}^k \in \mathbb{R}^{\overbrace{D \times D \times \ldots \times D}^{k \text{ times}}}$ is obtained as a k-fold outer product, denoted by $\otimes$, of the vector $\boldsymbol{\delta} \in \mathbb{R}^D$. For example,

$$\boldsymbol{\delta}^2 := \boldsymbol{\delta} \otimes \boldsymbol{\delta} = \boldsymbol{\delta}\boldsymbol{\delta}^\top , \quad \boldsymbol{\delta}^2[i,j] = \delta[i]\delta[j] \tag{5.153}$$

$$\boldsymbol{\delta}^3 := \boldsymbol{\delta} \otimes \boldsymbol{\delta} \otimes \boldsymbol{\delta} \,, \quad \boldsymbol{\delta}^3[i,j,k] = \delta[i]\delta[j]\delta[k] \,. \tag{5.154}$$

Figure 5.13 visualizes two such outer products. In general, we obtain the terms

$$D^k_{\boldsymbol{x}} f(\boldsymbol{x}_0)\boldsymbol{\delta}^k = \sum_{i_1=1}^{D} \cdots \sum_{i_k=1}^{D} D^k_{\boldsymbol{x}} f(\boldsymbol{x}_0)[i_1, \ldots, i_k]\delta[i_1] \cdots \delta[i_k] \tag{5.155}$$

in the Taylor series, where $D^k_{\boldsymbol{x}} f(\boldsymbol{x}_0)\boldsymbol{\delta}^k$ contains k-th order polynomials.

Now that we defined the Taylor series for vector fields, let us explicitly write down the first terms $D^k_{\boldsymbol{x}} f(\boldsymbol{x}_0)\boldsymbol{\delta}^k$ of the Taylor series expansion for $k = 0, \ldots, 3$ and $\boldsymbol{\delta} := \boldsymbol{x} - \boldsymbol{x}_0$:

$$k = 0 : D^0_{\boldsymbol{x}} f(\boldsymbol{x}_0)\boldsymbol{\delta}^0 = f(\boldsymbol{x}_0) \in \mathbb{R} \tag{5.156}$$

$$k = 1 : D^1_{\boldsymbol{x}} f(\boldsymbol{x}_0)\boldsymbol{\delta}^1 = \underbrace{\nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0)}_{1\times D} \underbrace{\boldsymbol{\delta}}_{D\times 1} = \sum_{i=1}^{D} \nabla_{\boldsymbol{x}} f(\boldsymbol{x}_0)[i]\delta[i] \in \mathbb{R} \tag{5.157}$$

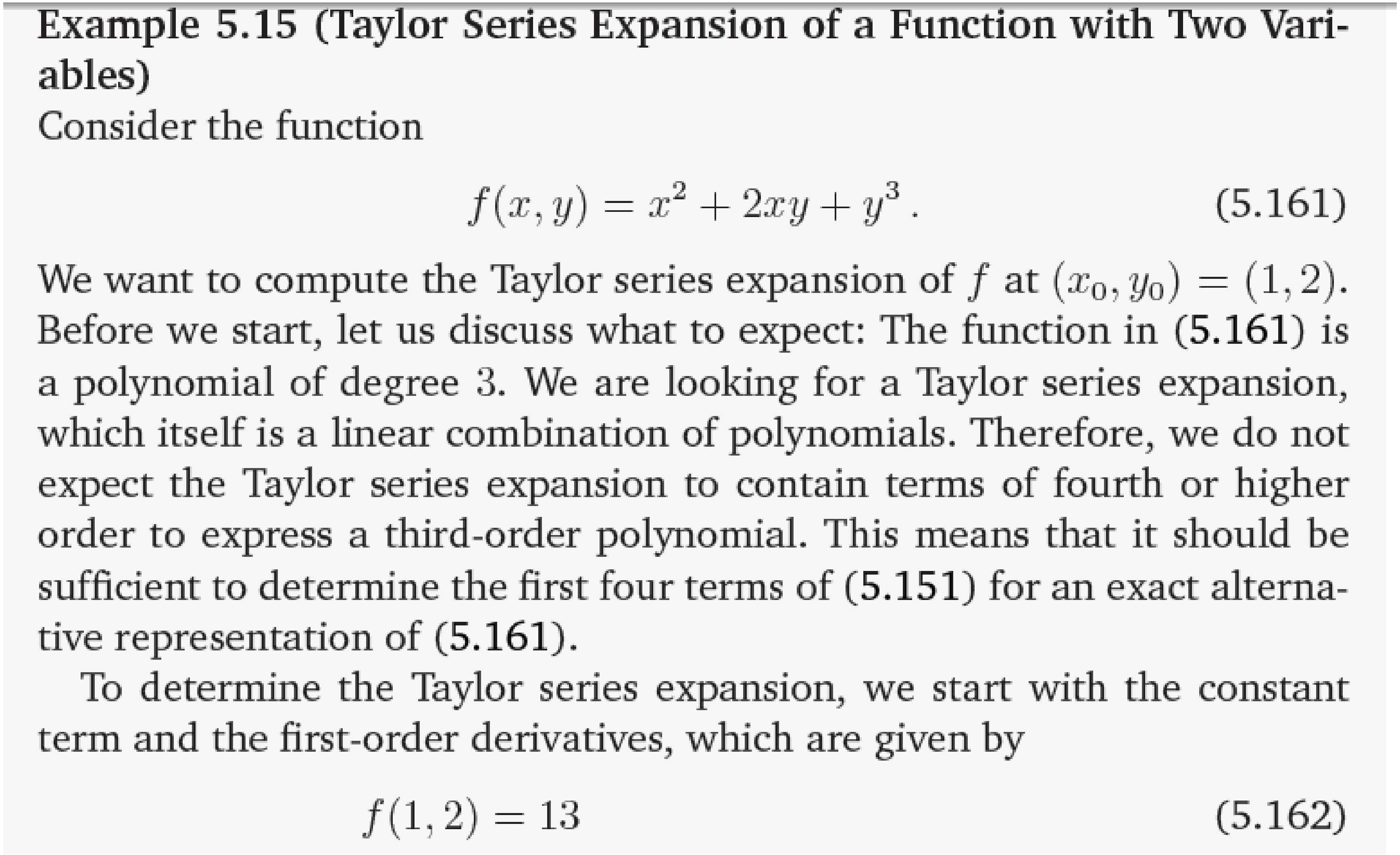
$$k = 2 : D^2_{\boldsymbol{x}} f(\boldsymbol{x}_0)\boldsymbol{\delta}^2 = \operatorname{tr}\big(\underbrace{\boldsymbol{H}(\boldsymbol{x}_0)}_{D\times D} \underbrace{\boldsymbol{\delta}}_{D\times 1} \underbrace{\boldsymbol{\delta}^\top}_{1\times D}\big) = \boldsymbol{\delta}^\top \boldsymbol{H}(\boldsymbol{x}_0)\boldsymbol{\delta} \tag{5.158}$$

$$= \sum_{i=1}^{D} \sum_{j=1}^{D} H[i,j]\delta[i]\delta[j] \in \mathbb{R} \tag{5.159}$$

$$k = 3 : D^3_{\boldsymbol{x}} f(\boldsymbol{x}_0)\boldsymbol{\delta}^3 = \sum_{i=1}^{D} \sum_{j=1}^{D} \sum_{k=1}^{D} D^3_{x} f(\boldsymbol{x}_0)[i,j,k]\delta[i]\delta[j]\delta[k] \in \mathbb{R} \tag{5.160}$$

Here, $\boldsymbol{H}(\boldsymbol{x}_0)$ is the Hessian of f evaluated at $\boldsymbol{x}_0$.

Example on Taylor Series



Example 5.15 (Taylor Series Expansion of a Function with Two Variables)

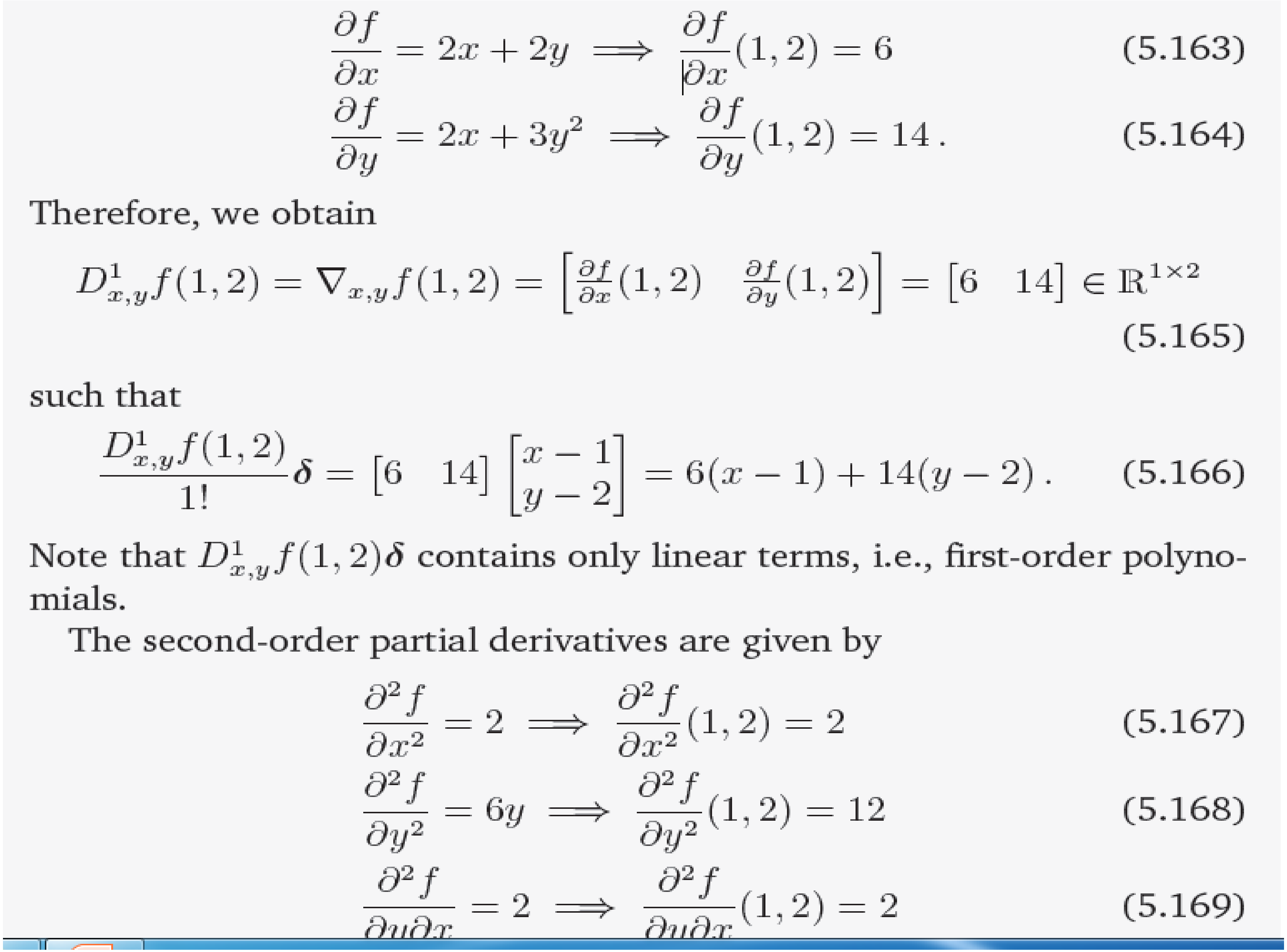
Consider the function

$$f(x, y) = x^2 + 2xy + y^3 \,. \tag{5.161}$$

We want to compute the Taylor series expansion of f at $(x_0, y_0) = (1, 2)$. Before we start, let us discuss what to expect: The function in (5.161) is a polynomial of degree 3. We are looking for a Taylor series expansion, which itself is a linear combination of polynomials. Therefore, we do not expect the Taylor series expansion to contain terms of fourth or higher order to express a third-order polynomial. This means that it should be sufficient to determine the first four terms of (5.151) for an exact alternative representation of (5.161).

To determine the Taylor series expansion, we start with the constant term and the first-order derivatives, which are given by

$$f(1, 2) = 13 \tag{5.162}$$



$$\frac{\partial f}{\partial x} = 2x + 2y \implies \frac{\partial f}{\partial x}(1,2) = 6 \tag{5.163}$$

$$\frac{\partial f}{\partial y} = 2x + 3y^2 \implies \frac{\partial f}{\partial y}(1,2) = 14\,. \tag{5.164}$$

Therefore, we obtain

$$D^1_{x,y}f(1,2) = \nabla_{x,y}f(1,2) = \begin{bmatrix}\frac{\partial f}{\partial x}(1,2) & \frac{\partial f}{\partial y}(1,2)\end{bmatrix} = \begin{bmatrix}6 & 14\end{bmatrix} \in \mathbb{R}^{1\times 2} \tag{5.165}$$

such that

$$\frac{D^1_{x,y}f(1,2)}{1!}\boldsymbol{\delta} = \begin{bmatrix}6 & 14\end{bmatrix}\begin{bmatrix}x-1\\ y-2\end{bmatrix} = 6(x-1) + 14(y-2)\,. \tag{5.166}$$

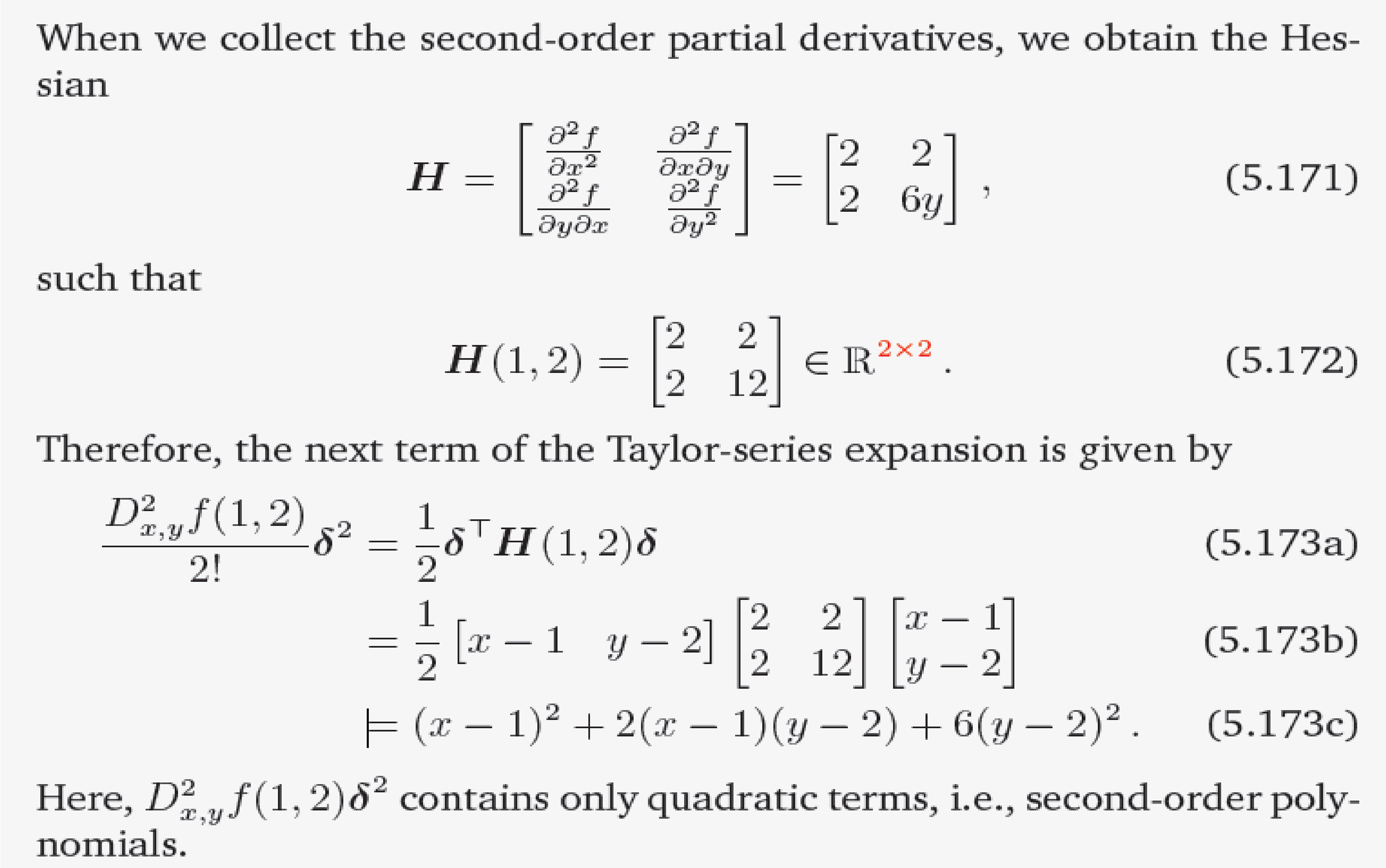
Note that $D^1_{x,y}f(1,2)\boldsymbol{\delta}$ contains only linear terms, i.e., first-order polynomials.

The second-order partial derivatives are given by

$$\frac{\partial^2 f}{\partial x^2} = 2 \implies \frac{\partial^2 f}{\partial x^2}(1,2) = 2 \tag{5.167}$$

$$\frac{\partial^2 f}{\partial y^2} = 6y \implies \frac{\partial^2 f}{\partial y^2}(1,2) = 12 \tag{5.168}$$

$$\frac{\partial^2 f}{\partial y \partial x} = 2 \implies \frac{\partial^2 f}{\partial y \partial x}(1,2) = 2 \tag{5.169}$$



When we collect the second-order partial derivatives, we obtain the Hessian

$$\boldsymbol{H} = \begin{bmatrix} \frac{\partial^2 f}{\partial x^2} & \frac{\partial^2 f}{\partial x \partial y} \\ \frac{\partial^2 f}{\partial y \partial x} & \frac{\partial^2 f}{\partial y^2} \end{bmatrix} = \begin{bmatrix} 2 & 2 \\ 2 & 6y \end{bmatrix}, \tag{5.171}$$

such that

$$\boldsymbol{H}(1,2) = \begin{bmatrix} 2 & 2 \\ 2 & 12 \end{bmatrix} \in \mathbb{R}^{2\times 2}. \tag{5.172}$$

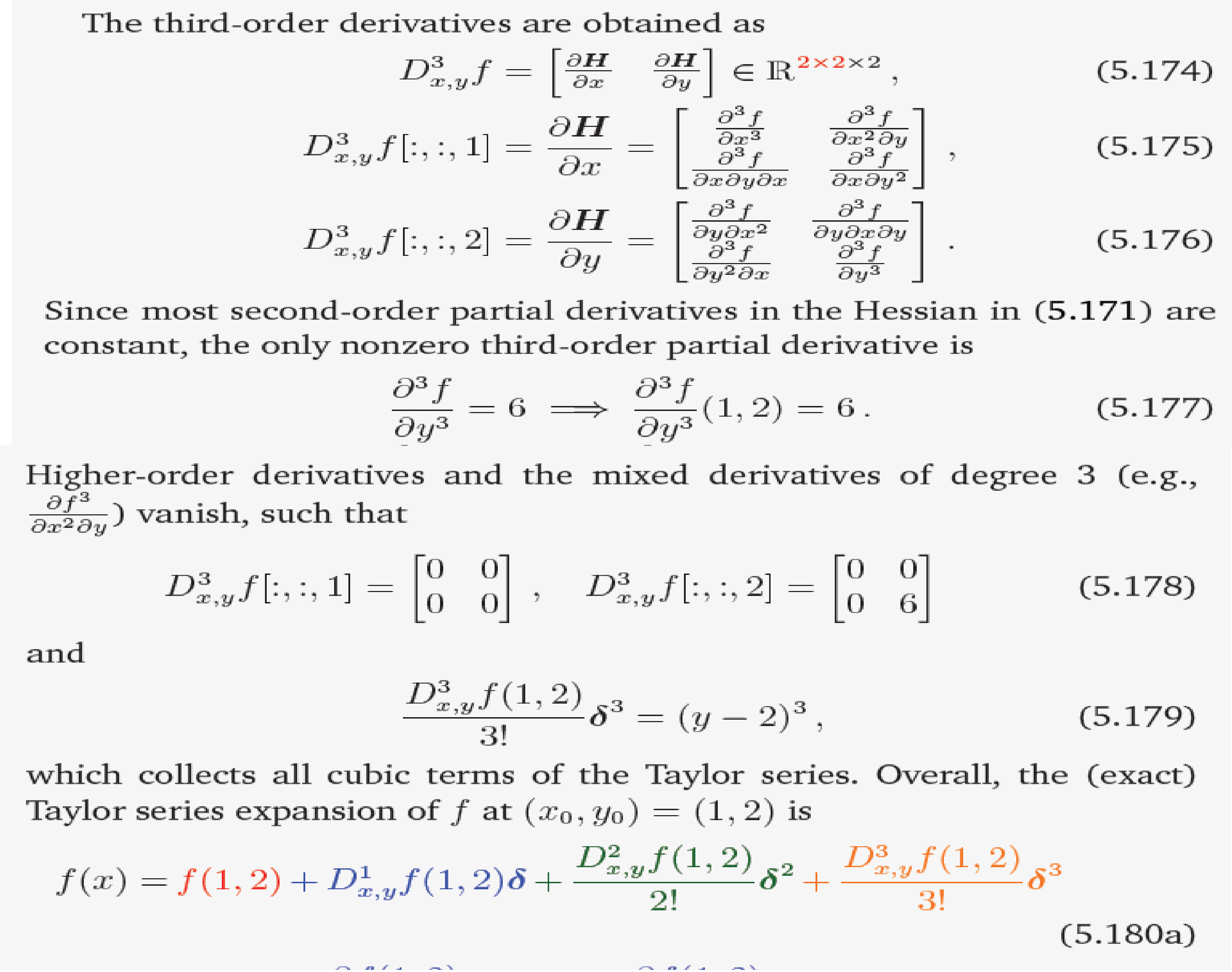
Therefore, the next term of the Taylor-series expansion is given by

$$\frac{D^2_{x,y} f(1,2)}{2!} \boldsymbol{\delta}^2 = \frac{1}{2} \boldsymbol{\delta}^\top \boldsymbol{H}(1,2) \boldsymbol{\delta} \tag{5.173a}$$

$$= \frac{1}{2} \begin{bmatrix} x-1 & y-2 \end{bmatrix} \begin{bmatrix} 2 & 2 \\ 2 & 12 \end{bmatrix} \begin{bmatrix} x-1 \\ y-2 \end{bmatrix} \tag{5.173b}$$

$$\models (x-1)^2 + 2(x-1)(y-2) + 6(y-2)^2. \tag{5.173c}$$

Here, $D^2_{x,y} f(1,2)\boldsymbol{\delta}^2$ contains only quadratic terms, i.e., second-order polynomials.



The third-order derivatives are obtained as

$$D^3_{x,y}f = \begin{bmatrix} \frac{\partial \boldsymbol{H}}{\partial x} & \frac{\partial \boldsymbol{H}}{\partial y} \end{bmatrix} \in \mathbb{R}^{2\times2\times2}\,, \tag{5.174}$$

$$D^3_{x,y}f[:,:,1] = \frac{\partial \boldsymbol{H}}{\partial x} = \begin{bmatrix} \frac{\partial^3 f}{\partial x^3} & \frac{\partial^3 f}{\partial x^2 \partial y} \\ \frac{\partial^3 f}{\partial x \partial y \partial x} & \frac{\partial^3 f}{\partial x \partial y^2} \end{bmatrix}\,, \tag{5.175}$$

$$D^3_{x,y}f[:,:,2] = \frac{\partial \boldsymbol{H}}{\partial y} = \begin{bmatrix} \frac{\partial^3 f}{\partial y \partial x^2} & \frac{\partial^3 f}{\partial y \partial x \partial y} \\ \frac{\partial^3 f}{\partial y^2 \partial x} & \frac{\partial^3 f}{\partial y^3} \end{bmatrix}\,. \tag{5.176}$$

Since most second-order partial derivatives in the Hessian in (5.171) are constant, the only nonzero third-order partial derivative is

$$\frac{\partial^3 f}{\partial y^3} = 6 \implies \frac{\partial^3 f}{\partial y^3}(1,2) = 6\,. \tag{5.177}$$

Higher-order derivatives and the mixed derivatives of degree 3 (e.g., $\frac{\partial f^3}{\partial x^2 \partial y}$) vanish, such that

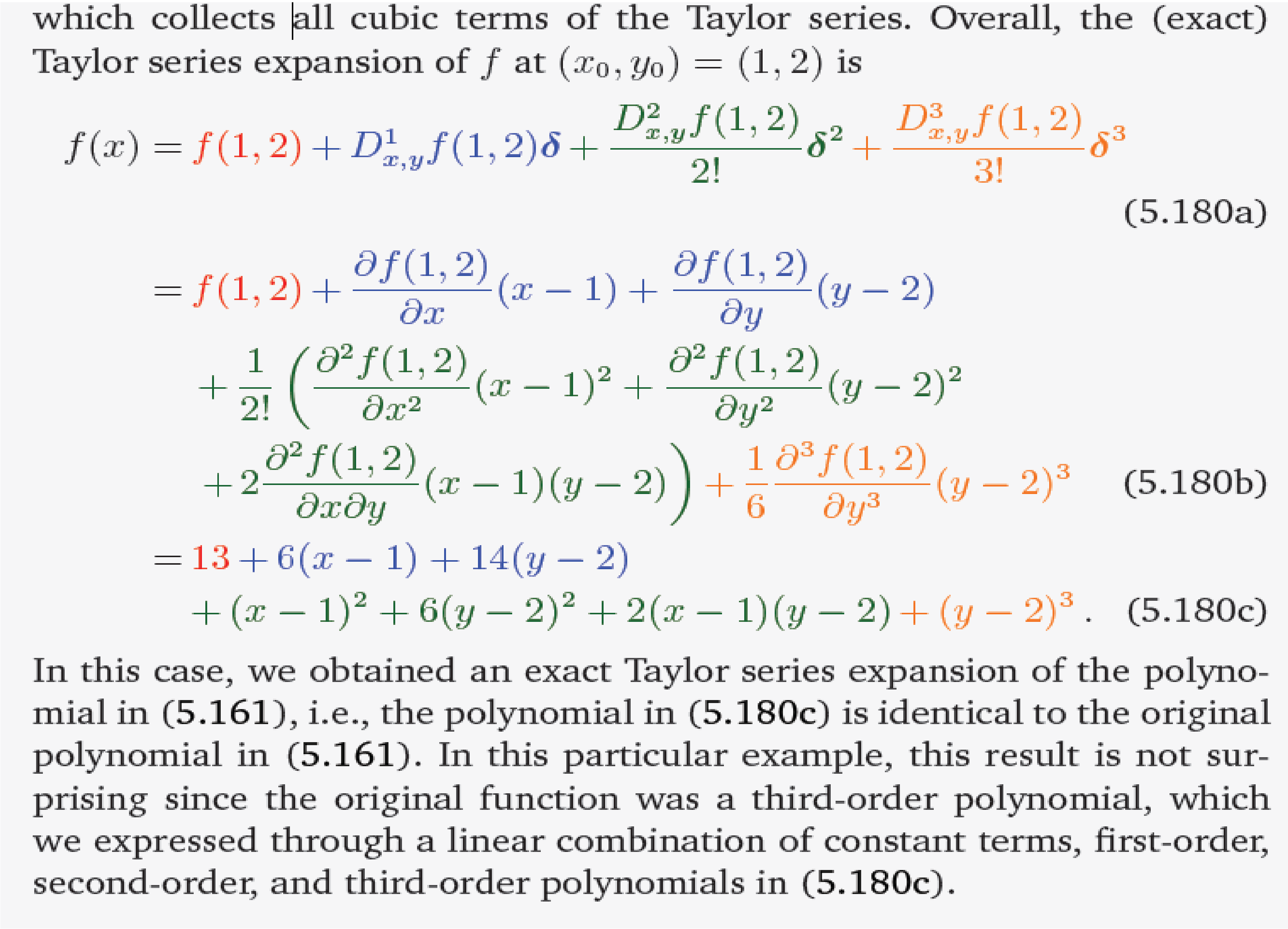
$$D^3_{x,y}f[:,:,1] = \begin{bmatrix} 0 & 0 \\ 0 & 0 \end{bmatrix}\,, \quad D^3_{x,y}f[:,:,2] = \begin{bmatrix} 0 & 0 \\ 0 & 6 \end{bmatrix} \tag{5.178}$$

and

$$\frac{D^3_{x,y}f(1,2)}{3!}\boldsymbol{\delta}^3 = (y-2)^3\,, \tag{5.179}$$

which collects all cubic terms of the Taylor series. Overall, the (exact) Taylor series expansion of f at $(x_0, y_0) = (1,2)$ is

$$f(x) = f(1,2) + D^1_{x,y}f(1,2)\boldsymbol{\delta} + \frac{D^2_{x,y}f(1,2)}{2!}\boldsymbol{\delta}^2 + \frac{D^3_{x,y}f(1,2)}{3!}\boldsymbol{\delta}^3 \tag{5.180a}$$



which collects all cubic terms of the Taylor series. Overall, the (exact) Taylor series expansion of f at $(x_0, y_0) = (1, 2)$ is

$$f(x) = f(1,2) + D^1_{x,y}f(1,2)\boldsymbol{\delta} + \frac{D^2_{x,y}f(1,2)}{2!}\boldsymbol{\delta}^2 + \frac{D^3_{x,y}f(1,2)}{3!}\boldsymbol{\delta}^3 \tag{5.180a}$$

$$\begin{aligned} &= f(1,2) + \frac{\partial f(1,2)}{\partial x}(x-1) + \frac{\partial f(1,2)}{\partial y}(y-2) \\ &\quad + \frac{1}{2!}\left(\frac{\partial^2 f(1,2)}{\partial x^2}(x-1)^2 + \frac{\partial^2 f(1,2)}{\partial y^2}(y-2)^2\right. \\ &\quad \left. + 2\frac{\partial^2 f(1,2)}{\partial x \partial y}(x-1)(y-2)\right) + \frac{1}{6}\frac{\partial^3 f(1,2)}{\partial y^3}(y-2)^3 \end{aligned} \tag{5.180b}$$

$$\begin{aligned} &= 13 + 6(x-1) + 14(y-2) \\ &\quad + (x-1)^2 + 6(y-2)^2 + 2(x-1)(y-2) + (y-2)^3 \,. \end{aligned} \tag{5.180c}$$

In this case, we obtained an exact Taylor series expansion of the polynomial in (5.161), i.e., the polynomial in (5.180c) is identical to the original polynomial in (5.161). In this particular example, this result is not surprising since the original function was a third-order polynomial, which we expressed through a linear combination of constant terms, first-order, second-order, and third-order polynomials in (5.180c).