Mathematical Appendix

This section introduces some key mathematical concepts required to understand the material in these notes, along with the corresponding mathematical notation. After introducing sets and some elementary notation, it discusses functions, along with differentiation, integration and optimization. Then, it introduces matrix-notation and discusses some basic definitions, operations and results in the context of matrices (and vectors).

Long formulas and algebraic manipulations might appear off-putting for the reader not experienced with linear algebra. However, all of the manipulations in these notes make use of a small set of simple results. After becoming familiar with them and with some training, some of the longer manipulations in these notes might still be cumbersome, but not difficult. Similarly, readers without a quantitative background might shy away from mathematical notation. However, once one gets used to it, one starts to appreciate its advantages. First, it is highly efficient as it can write with a few characters what in words would take several lines. Second, it is very precise, whereas the exact meaning of verbal statements can be ambiguous.

M1 Sets & Elementary Notation

Sets appear in all areas of mathematics, statistics and any subject that builds on them, like econometrics. A set is a collection of elements. For example, $\{1, 2, 3\}$ is a set with three elements: the three integers 1, 2 and 3. In principle, the elements of a set could be anything. For example, we could define the sets $\{A, B\}$ or $\{bottle, book\}$ or $\{3, computer, \gamma\}$. Sometimes, the elements in a set are themselves sets or, more generally, a list of elements (possibly arranged in a specific order): $\{\{1, 1\}, \{1, 2\}, \{2, 1\}\}$ or $\{(1, 1), (1, 2), (2, 1)\}$.

A commonly used set is the set of natural numbers, $\{1, 2, 3, ...\}$, denoted by $\mathbb{N}$. The set of integers, $\{..., -3, -2, -1, 0, 1, 2, 3, ...\}$, is denoted by $\mathbb{Z}$. The set of rational numbers, i.e. numbers that can be written as a fraction involving two integers, is denoted by $\mathbb{Q}$. Finally, and most importantly, the set of real numbers is denoted by $\mathbb{R}$. If we consider only the

non-negative real numbers, we write $\mathbb{R}_+$. If we consider only positive real numbers (i.e. we exclude zero), we write $\mathbb{R}_{++}$.

While there is a whole mathematical field of set theory, we are more concerned with the notion of a set and the elementary notation around it rather than all the operations that one can do with sets.[1] The character $\in$ denotes "(is) element of". For example, $3 \in \mathbb{N}$. Analogously, $\notin$ is the negation of $\in$, i.e. it denotes "(is) not element of": $3.5 \notin \mathbb{N}$. If we want to denote a generic element of a set, we write it as a variable, i.e. we use a letter for it: $x \in \{1, 2, 3\}$ denotes an (any) element in the set $\{1, 2, 3\}$. With this notation, we can do operations or define statements for all elements of a set. For example, to say "each element of the set $\{1, 2, 3\}$ is smaller than 4", we write

$$x < 4 \ \forall \ x \in \{1, 2, 3\} \ ,$$

whereby the character $\forall$ means "for each". Another useful character is $\exists$, which means "there exists". For example, we could write

$$\exists \, x \in \{1, 2, 3\} \text{ s.t. } x < 3 \ ,$$

where s.t. means "such that", i.e. "there exists an element in the set $\{1, 2, 3\}$ which is smaller than 3". Similarly, we can write $\nexists \, x \in \{1, 2, 3\}$ s.t. $x > 4$. Some statements can be written in several ways. For example, $x > 0$ and $x \in \mathbb{R}_{++}$ mean the same thing.

Note that $\mathbb{N}$ is a subset of $\mathbb{Z}$, i.e. all the elements of $\mathbb{N}$ are also in $\mathbb{Z}$, that is,

$$x \in \mathbb{Z} \ \forall \ x \in \mathbb{N} \ .$$

We write $\mathbb{N} \subseteq \mathbb{Z}$.[2] Similarly, $\mathbb{N}$, $\mathbb{Z}$ and $\mathbb{Q}$ are all subsets of $\mathbb{R}$.

We can write the above statement $x \in \mathbb{Z} \ \forall \ x \in \mathbb{N}$ also as

$$x \in \mathbb{N} \Rightarrow x \in \mathbb{Z} \ ,$$

whereby the character $\Rightarrow$ means "implies". We can read the statement $A \Rightarrow B$ also as "if statement A holds then statement B holds". We say that statement A is a sufficient condition for statement B. However, A is not necessary for B; we could have statement B even without statement A. For example, saying "When I go for a run, I wear a red shirt"

[1] The Appendix to Chapter 1 introduces some elementary operations – like unions, intersections, complements – necessary to understand elementary probability theory.

[2] In fact, we have $\mathbb{N} \subset \mathbb{Z}$, as there are elements in $\mathbb{Z}$ that are not in $\mathbb{N}$, i.e. we write $A \subset B$ if $A \subseteq B$ but not $B \subseteq A$.

(run $\Rightarrow$ red shirt) does not imply that whenever I wear a red shirt I must be running. To know whether I wear a red shirt, it would be sufficient to know that I went for a run, but it is not necessary; I can wear a red shirt even when not running. In the above example of integers and natural numbers, we can have $x \in \mathbb{Z}$ even for $x \notin \mathbb{N}$.[3] If for two statements A and B we have $A \Rightarrow B$ as well as $B \Rightarrow B$, then we have a both-sided implication and we write $A \Leftrightarrow B$. This reads as "statement A holds iff (if and only if) statement B holds". In other words, statement A is a necessary and sufficient condition for statement B (and vice versa), i.e. the two statements are equivalent.

We can define a set based on another set. For example, another way to denote the set $\{1, 2, 3\}$ is to write

$$\{x \in \mathbb{N} : x < 4\} \quad \text{or} \quad \{x \in \mathbb{N} \text{ s.t. } x < 4\} .$$

This is read as "all elements x in $\mathbb{N}$ s.t. x is smaller than 4". This is useful if we need to define some more involved set, like all natural numbers divisible by three or all (positive and negative) integers divisible by three, which is, respectively,

$$\{x \in \mathbb{N} : x/3 \in \mathbb{N}\} \quad \text{and} \quad \{x \in \mathbb{Z} : x/3 \in \mathbb{N} \text{ or } -x/3 \in \mathbb{N}\} .$$

The latter is read as "all integers x s.t. $x/3$ is a natural number (i.e. x is positive and divisible by three) or $-x/3$ is a natural number (i.e. x is negative and divisible by three)". To get the set of all pairs $(1, 1), (1, 2), (1, 3), ...$, we can write $\{(1, x) : x \in \mathbb{N}\}$. To get all the intervals $[x, x + 5)$, whereby x is a positive real number, we write $\{[x, x + 5) : x \in \mathbb{R}_{++}\}$. In this context, it is useful to define the empty set, denoted by $\emptyset$. For example, we have

$$\{x \in \mathbb{N} : x < 0\} = \emptyset .$$

This is because $\nexists\, x \in \mathbb{N}$ s.t. $x < 0$.

When dealing with sets whose elements are lists (like "pairs", vectors or matrices), it is useful to know the concept of a Cartesian product. For example, the Cartesian product of $\{1, 2\}$ and $\{a, b\}$ – denoted by $\{1, 2\} \times \{a, b\}$ – is the set $\{(1, a), (1, b), (2, a), (2, b)\}$ or $\{\{a, b\}, \{1, b\}, \{2, a\}, \{2, b\}\}$, depending on one's preferred notation and the context.[4] We can take the Cartesian product of a set with itself: e.g. $\mathbb{N} \times \mathbb{N} = \mathbb{N}^2$, which is the set with

[3] Note that $A \Rightarrow B$ is equivalent to $\not B \Rightarrow \not A$, where $\not$ denotes a negation and is read as "not". For example, saying "When I go for a run, I wear a red shirt" (run $\Rightarrow$ red shirt) is equivalent to saying "If I do not wear a red shirt, I cannot be running" (not red shirt $\Rightarrow$ not run) (or we would rather say "I never go for a run without a red shirt").

[4] Note that the ordering of elements in a set is irrelevant, in principle (though sometimes there is a natural ordering which it is reasonable to adhere to).

elements (x, y) whereby x and y are both natural numbers:

$$\mathbb{N}^2 = \{(x, y) : x, y \in \mathbb{N}\} .$$

As a result, the set of all three-dimensional vectors (i.e. just all ordered lists of three elements; see below) with elements that are real numbers is denoted by $\mathbb{R}^3$. We speak of "real-valued" vectors. The set of all real-valued vectors of length n is $\mathbb{R}^n$.

M2 Functions

A function is a mapping from one set to another. A typical example is a function that takes a real number as the argument and returns another real number as the output, i.e. it maps one real number into another. For example, $f(x) = x^2$ maps $x \in \mathbb{R}$ into $x^2 \in \mathbb{R}$. To refer generically to this function, we write f or $f(x)$ (or $f(z)$ for that matter, the variable we assign to the argument is irrelevant, it does not change the function). To refer to the function evaluated at a specific point, we write e.g. $f(3)$, which gives 9 in this case, or we write more generically $f(x^*)$ for some (particular) $x^* \in \mathbb{R}$.[5]

The function

$$g(x) = \begin{cases} 0 & \text{if } x < 1/6 \\ 1 & \text{otherwise (i.e. } x \geq 1/6) \end{cases}$$

maps $\mathbb{R}$ into $\{0, 1\}$. We call the first set the "domain" of our function g and the second set its "codomain". A short way to write this same function is $g(x) = \mathbf{1}\{x < 1/6\}$, whereby $\mathbf{1}\{\cdot\}$ is the indicator function (or indicator-operator). It returns a one if the condition inside the brackets is true and a zero otherwise. To get a function that returns a 2 if $x < 1/6$ and a 5 otherwise, we can write $f(x) = 2 + 3\,\mathbf{1}\{x < 1/6\}$. An example of a function that maps natural numbers into natural numbers is the factorial: $f(x) = x!$. It is not defined for an argument x which is not a natural number, e.g. $x = 2.4$. Similarly, the function $f(x) = 1/x$ maps real numbers into real numbers, but it is not defined for $x = 0$. In mathematical notation, it maps $\mathbb{R}\backslash\{0\}$ (all the real numbers except zero) into $\mathbb{R}$.

To generically denote some function that maps a set A into a set B, we write $f : A \to B$. This is useful in mathematical theorems, as one might want to posit that something holds for all real-valued functions, i.e. $\forall\, f : \mathbb{R} \to \mathbb{R}$. It can also be useful if a specific function is

[5]Because we could in principle use any variable to denote the value of the function evaluated at this variable, it is preferrable to write f rather than $f(x)$ when referring to the function itself. Nevertheless, people often write $f(x)$ for the function itself and use additional characters like asterisks or hats or sub- and superscripts to denote a specific point at which the function is evaluated.

too complicated to write down, but it is implicitly defined by some procedure and one wants to specify its domain and codomain.[6]

Important functions to be familiar with include the exponential function, $exp\{x\}$ or e^x, and the natural logarithm, $\ln x$, which in these notes is also written as $\log x$ (as logarithms with other bases are not used). The former maps $\mathbb{R}$ into $\mathbb{R}_{++}$, while the latter maps $\mathbb{R}_{++}$ into $\mathbb{R}$. It holds that

$$e^x e^y = e^{x+y} \ , \quad \log xy = \log x + \log y \quad \text{and} \quad \log x^b = b \log x \ .$$

Also, $\log exp\{x\} = exp\{\log x\} = x$. A linear function takes the form $f(x) = a + bx$ for some $a, b, \in \mathbb{R}$, while quadratic functions are $f(x) = a + bx + cx^2$ for $a, b, c \in \mathbb{R}$. Thereby, a, b, c are parameters of the function, to be distinguished from its argument, x. Note that we can define the set of quadratic functions as $\{f : f(x) = a + bx + cx^2 \ , \ a, b, c \in \mathbb{R}, x \in \mathbb{R}\}$.

A function can be inverted. For example, if $f(x) = 3x$, then $f^{-1}(x) = x/3$. Similarly, for $f(x) = x^2$, we have $f^{-1}(x) = \pm\sqrt{x}$. We call f^{-1} the inverse or inverse-image of f, as it undoes the mapping performed by f. Note that the natural logarithm and the exponential function are inverse-images of one another. A function is said to be injective or "one-to-one" if for any single, unique element in the domain (argument) it returns a different, unique element in the codomain (output).[7] For example, $f(x) = 3x$ is an injective function as for different xs it returns different outputs $3x$. If the reverse is also true, then we speak of a bijective function. For example, $f(x) = 3x$ is bijective, as for any number $3x$, we can find the unique x that gives this number by dividing by three.[8] In contrast, $f(x) = x^2$ is only injective, not bijective, as $f(x) = 9$ could mean that $x = 3$ or $x = -3$.[9] Bijectivity is a useful property in many different contexts. It implies that rather than working with some original variable x, we can define a new variable that is a linear function of x, and, once we are done with our math, we can uniquely find x based on our new variable. Note that the exponential function and the natural logarithm as well as linear functions are bijective.

A function is said to be continuous if it has no jumps, i.e. no discontinuities, otherwise it is discontinuous. Essentially, a continuous function is one that can be drawn on a graph

[6] In principle, we could also define a function like $f(x) = x^2$ to take only arguments in $\mathbb{N}$. However, this is rarely useful. Also, functions can map more abstract sets into one another, e.g. $f : \{\text{dry, rain, snow}\} \to \{\text{picnic possible, picnic not possible}\}$.

[7] In short, a function f is injective if $f(a) = f(b) \Rightarrow a = b$, as this is equivalent to saying $a \neq b \Rightarrow f(a) \neq f(b)$.

[8] Note that a function f being bijective is equivalent to f and f^{-1} both being injective.

[9] A function can also be surjective (i.e. only f^{-1} is injective, not f), and it can be neither of these three definitions.

without lifting the pencil off the paper. Mathematically, a function f is continuous if

$$\lim_{x \to c} f(x) = f(c) \; \forall \; c.^{10}$$

The function $f(x) = x^2$ is continuous. So is the function $f(x) = |x|$, despite its spike at zero. The function $f(x) = \mathbf{1}\{x < 1/6\}$ is not continuous due to a discontinuity at $x = 1/6$. The function $f(x) = 1/x$ has a discontinuity at $x = 0$, but it is continuous on its domain, $\mathbb{R}\backslash\{0\}$, as the latter does not include zero.

The above examples all discussed functions that take a single, scalar-argument and return a single, scalar-output. A scalar is a single number, most generally a number in $\mathbb{R}$.[11] In contrast, vectors and matrices are ordered lists/collections of multiple scalars (numbers) (see below). Functions can have multiple (scalar-) arguments and/or have multiple (scalar-) outputs, i.e. they can take vectors or matrices as arguments and/or return vectors or matrices as output. For example, the function $f(x, y) = x^3 + e^y$ takes two arguments and returns a real number, i.e. it maps $\mathbb{R}^2$ into $\mathbb{R}$. The function $f(x, y) = x^3 + \log y$ maps $\mathbb{R} \times \mathbb{R}_{++}$ into $\mathbb{R}$.[12] Rather than distinguishing the individual, scalar-arguments x and y, we might also just write $f(v)$, where $v \in \mathbb{R}^2$ (or $v \in \mathbb{R} \times \mathbb{R}_{++}$) is a vector with two elements. These cases are discussed in more detail below, once matrix notation is introduced. If not otherwise specified, in the following we deal with functions with scalar-arguments and -outputs.

M2.1 Differentiation

The first-order derivative of a function $f : \mathbb{R} \to \mathbb{R}$ is defined as

$$f'(x) = \lim_{h \to 0} \frac{f(x + h) - f(x)}{h} \; .$$

On a graph, it is the slope of f. Note that f' is itself also a function. Just as the function f, it too can be evaluated at a specific point, say x^*. We write the derivative f' also as $f^{(1)}$ (or $f^{(1)}(x)$) or $\frac{\partial f(x)}{\partial x}$.[13] To then evaluate it at a point x^*, we write $f'(x^*)$, $f^{(1)}(x^*)$ or $\frac{\partial f(x)}{\partial x}\Big|_{x=x^*}$. Depending on the context, one of these three ways can be neater to use than the others.

It is useful to memorize the derivatives of some functions (expressions). Let $a, c \in \mathbb{R}$ and

[10] Thereby, x and c have to be points in the domain of f. More precisely, we look at functions $f : D \to \mathbb{R}$, whereby $D \subseteq \mathbb{R}$ is a subset of $\mathbb{R}$, and we require $x, c \in D$, i.e. x approaches c through the domain of f.

[11] I write most generally because any number in $\mathbb{N}$, $\mathbb{Z}$ or $\mathbb{Q}$ is also in $\mathbb{R}$, i.e. the former three sets are all subsets of $\mathbb{R}$.

[12] Note that log can only be applied to positive numbers.

[13] Equivalently, we can also write $\frac{\partial}{\partial x} f(x)$.

$b \in \mathbb{R}_+$. We have

$$\begin{aligned}\frac{\partial}{\partial x}\left[c + ax^b\right] &= abx^{b-1} \ , \\ \frac{\partial}{\partial x}\log x &= \frac{1}{x} \ , \\ \frac{\partial}{\partial x} exp\{x\} &= \exp\{x\} \ .\end{aligned}$$

For example, we have $\frac{\partial}{\partial x}2x^3 = 6x$. Also, there are a couple of rules that are useful to remember:

$$\begin{aligned}\frac{\partial}{\partial x}af(x) &= af'(x) \ , \\ \frac{\partial}{\partial x}\left[f(x) + g(x)\right] &= f'(x) + g'(x) \ , \\ \frac{\partial}{\partial x}f(x)g(x) &= f'(x)g(x) + f(x)g'(x) \ , \text{[product rule]} \\ \frac{\partial}{\partial x}\frac{f(x)}{g(x)} &= \frac{f'(x)g(x) - f(x)g'(x)}{g(x)^2} \ .\end{aligned}$$

As a result, we have, for example $\frac{\partial}{\partial x}\left[\log x - 3x^2\right] = 1/x - 6x$, $\frac{\partial}{\partial x}\log(x)(3x^2) = \frac{1}{x}3x^2 + \log(x)6x = 3x(1+2\log(x))$, $\frac{\partial}{\partial x}\frac{1}{x} = -\frac{1}{x^2}$, or $\frac{\partial}{\partial x}\frac{1}{3x^2} = -\frac{6x}{9x^4} = -\frac{2}{3x^3}$. Finally, to take derivatives of functions of functions, we use the rule

$$\frac{\partial}{\partial x}f(g(x)) = f'(g(x))g'(x) = \left.\frac{\partial f(z)}{\partial z}\right|_{z=g(x)} g'(x) \ ,$$

i.e. to take the derivative of $f(g(x))$, we first take the derivative of f and evaluate it at $g(x)$, and we then multiply this by the derivative of g. This gives, for example, $\frac{\partial}{\partial x}\log(3x^2) = \frac{1}{3x^2}6x = \frac{2}{x}$.

A function f is said to be differentiable if its derivative is defined everywhere, i.e. for all points in its domain. Examples of non-differentiable functions are $f(x) = |x|$ and $f(x) = \mathbf{1}\{x < 1/6\}$. The former is not differentiable at $x = 0$, the latter at $x = 1/6$.[14] [15] A function f is said to be continuously differentiable if it is differentiable (i.e. its derivative exists everywhere) and its derivative is a continuous function. The function $f(x) = |x|$ is

[14] Often, rather than saying a function is not differentiable, one says it is not differentialble everywhere or one specifies the point where it is not differentiable.

[15] Note that these two examples show that differentiability and continuity are not the same thing, as neither function is differentiable, but the former function is continuous, the latter is not.

differentiable but not continuously differentiable because

$$f'(x) = \begin{cases} 1 & \text{if } x > 0 \\ -1 & \text{if } x \leq 0 \end{cases}$$

is not a differentiable function.

We can also take higher-order derivatives of a function, which involves nothing else than repeatedly taking derivatives, i.e. taking derivatives of derivatives. For example, the second-order derivative of a function f is

$$f''(x) = \frac{\partial f'(x)}{\partial x} \ .$$

We also write $f^{(2)}(x)$ or $\frac{\partial^2 f(x)}{\partial^2 x}$.[16] For example, we have $\frac{\partial^2}{\partial^2 x} 2x^3 = 6$ since the derivative of $6x$ is 6. The kth order derivative of f is denoted as $f^{(k)}$ or $\frac{\partial^k f(x)}{\partial^k x}$.

Note that for functions with multiple (scalar-) arguments, we can take derivatives w.r.t. (with respect to) different arguments. For example, for $f(x, y) = x^3 + e^y$, we have $\frac{\partial f}{\partial x} = 3x^2$ and $\frac{\partial f}{\partial y} = e^y$.

M2.2 Integration

Differentiation is one of the two fundamental operations of (the mathematical branch of) calculus. The other is integration. For a function defined on a discrete domain (i.e. a domain which is a set with countably many elements), like $\{1, 2, ..., n\}$, we could simply sum up the function evaluated at different all the different x in the domain:

$$\sum_{x \in \{1,2,...,n\}} f(x) = \sum_{x=1}^{n} f(x) \ .$$

For example, for $f(x) = x^2$, we get $1^2 + 2^2 + ... + n^2 = n(n+1)(2n+1)/6$. In contrast, we cannot do so for a function defined on a continuous domain, i.e. a set which has uncountably many elements, like $\mathbb{R}$. For such functions, we can compute the integral, i.e. the area under the curve, e.g. between the points $x = a$ and $x = b$:

$$\int_a^b f(x)dx \ .$$

[16] This turns out to be the same as $\frac{\partial^2 f(x)}{\partial x^2}$.

The integral of a continuous function is the counterpart of a sum for discrete functions. It turns out that

$$\int_a^b g'(x)dx = g(b) - g(a) \ ,$$

i.e. to compute the area under a function f between the points a and b, we need to find the function g which, when differentiated, gives the function f, and evaluate it at b and at a. For example, we get

$$\int_1^2 3x^2 dx = (x^3 + c)\big|_{x=2} - (x^3 + c)\big|_{x=1} = 8 - 1 = 7 \ .$$

Note that the constant c cancels out. We write $g(b) - g(a)$ more compactly as $[g(x)]_a^b$. Sometimes we leave out the bounds of the integration because we want to integrate a function over its whole domain. For example, if the domain is $\mathbb{R}$, rather than writing $\int_{-\infty}^{\infty} f(x)dx = \int_{\mathbb{R}} f(x)dx$, we could simply write $\int f(x)dx$. We also write the integral without bounds if we want to compute the function which, when differentiated, gives our original function f (often referred to as "the integral of f"). For example, we have $\int 3x^2 dx = x^3$.

A useful formula for computing integrals is integration by parts:

$$\int_a^b f(x)g'(x)dx = [f(x)g(x)]_a^b - \int_a^b f'(x)g(x)dx \ .$$

It is useful when we can split up the function whose integral we want to compute into a function f for which we can easily compute its derivative, f', and a function g' for which we can easily compute its integral, g.

M2.3 Optimization

Sometimes, we want to find the optimum of a function, i.e. its maximum or minimum (highest and lowest point, respectively, when dealing with a function with scalar-arguments and -outputs and plotting it on a graph). The operation of finding the minimum or maximum is written as

$$\min_x \ f(x) \quad \text{or} \quad \max_x \ f(x) \ .$$

The subscript x appears because for functions that have several arguments, it is important to specify w.r.t. which argument we are performing the optimization. The point at which the optimum is reached is denoted as

$$x^* = \arg\min_x \ f(x) \quad \text{or} \quad x^* = \arg\max_x \ f(x) \ .$$

This reads as "x^* is the argument that minimizes (maximizes) the function f". Because the opimum might not be unique, we should write more precisely $x^* \in \arg\min_x \ f(x)$, i.e. "$x^*$ is an argument that minimizes (maximizes) the function f", i.e. it is in the set of values that minimize (maximize) the function f. Similarly, an optimization problem might also have no solution, i.e. the set of optimizers is the empty set. The value of the function evaluated at the optimum (i.e. at one of the optimizers) is then written in one of the following ways:

$$f(x^*) = \min_x \ f(x) \quad \text{or} \quad f(x^*) = \max_x \ f(x) \ .$$

Sometimes it is also important to specify the domain over which we are optimizing a function f, i.e. the set of values that x can take on. For example,

$$\max_{x \in \mathbb{R}} x^2 \quad \text{is different than} \quad \max_{x \in [0,2]} x^2 \ ,$$

because the former optimization problem has no solution, while the second gives $x^* = 2$. If an optimizer is not at the bounds of the domain (e.g. not 0 or 2 in the above example), we speak of an interior optimum and interior optimizer. For example,

$$\min_{x \in \mathbb{R}} x^2 \quad \text{and} \quad \min_{x \in [-1,2]} x^2$$

both have the interior solution $x^* = 0$.

An interior optimum of a differentiable function can be found by setting the first-order derivative to zero. This is referred to as the first-order condition (FOC). The intuition is that the slope at the optimum has to be zero. For example, the minimium of $f(x) = 2 + 3x^2$ satisfies $6x = 0$, i.e. we have the minimum $x^* = 0$, at which f takes on the value $f(0) = 2$, i.e.

$$0 = \arg\min_x 2 + 3x^2 \ , \quad \min_x 2 + 3x^2 = f(0) = 2 \ .$$

To be sure that we found a minimum (maximum), we also need to verify that the second-order derivative is positive (negative) at the optimum.[17] [18] In our example, we get $f''(x) = 6 > 0 \ \forall \ x$, so we can be sure that $x^* = 0$ is a minimum. However, we oftentimes – as here – deal with globally convex (concave) functions, which – for differentiable functions – means that the second-order derivatives are positive (negative) everywhere, respectively.

[17] The intuition is that at the minimum, the slope changes from negative to positive, i.e. the slope of the slope is positive, and analogously it is negative for a maximum.

[18] However, note that for some function, the second derivative does not provide conclusive evidence on whether we found a minimum (maximum). For example, for $f(x) = x^4$ we have $f''(x) = 12x^2$ which is zero at the minimum $x^* = 0$.

For non-differentiable functions, we have to find the minimum (maximum) by showing with some other argument that $f(x^*) < (>) f(x) \; \forall \; x \neq x^*$. For example, $f(x) = 2 + 3|x|$ is also minimized at $x = 0$ as can be seen by plotting this function.[19]

Note that if x^* is the minimizer (maximizer) of $f(x)$, then it is also the minimizer (maximizer) of a monotonic transformation of f, like $a + bf(x)$ for $b > 0$ or $\log f(x)$, and vice versa.[20] This is useful as f might involve a big product or an exponential function, and the log of a product gives the sum of the logs and $\log \exp\{x\} = x$ (see above), for which it is easier to compute derivatives.

M3 Vectors & Matrices

In contrast to a scalar, which is just a single number, vectors and matrices are ordered lists of several scalars, arranged in a rectangular way. As an example, consider the 2×1 vector v and the 2×3 matrix M:

$$v = \begin{bmatrix} 3 \\ 1 \end{bmatrix} \; , \quad M = \begin{bmatrix} 2 & 2 & 7 \\ 4 & 9 & 3 \end{bmatrix} \; .$$

Note that a vector and a scalar are just special types of a matrix; the former is a $k \times 1$ matrix, the latter is a 1×1 matrix.

M3.1 Elementary Operations & Definitions

A vector is a column-vector if it has just one column (i.e. it is arranged vertically), whereas it is a row-vector if it has just one row (i.e. it is arranged horizontally). We will by default always use column-vectors. However, it is easy to transform them into row-vectors by transposing them. The transpose is the operation that flips the elements of a vector or matrix so that we get

$$v' = \begin{bmatrix} 3 & 1 \end{bmatrix} \; , \quad M' = \begin{bmatrix} 2 & 4 \\ 2 & 9 \\ 7 & 3 \end{bmatrix} \; .$$

Transposing a scalar s (i.e. a 1×1 matrix) does not change it: $s' = s$.

We can add together two matrices (vectors), provided that they have the same dimensions.

[19] See the comment on monotonic transformations below: $x = 0$ is the minimizer because $a + b|x|$ is a monotonic transformation of $|x|$, and we have $|x| \geq 0 \; \forall \; x$ and $|x| = 0$ only at $x = 0$.

[20] i.e. $x^* = \arg\min_x f(x) \Leftrightarrow x^* = \arg\min_x g(f(x))$, where g is some monotonic transformation of f.

For example, we get

$$\begin{bmatrix}3\\1\end{bmatrix}+\begin{bmatrix}1\\-2\end{bmatrix}=\begin{bmatrix}4\\-1\end{bmatrix}\ ,\quad \begin{bmatrix}2&2&7\\4&9&3\end{bmatrix}+\begin{bmatrix}1&-3&2\\5&-3&6\end{bmatrix}=\begin{bmatrix}3&-1&9\\9&6&9\end{bmatrix}\ .$$

The new matrix (vector) has the same dimensions as the two matrices (vectors) that are added together, and its elements are simply the element-wise additions of the elements of the two matrices (vectors). Multiplying a matrix (vector) by a scalar just involves multiplying each element by this scalar:

$$s\begin{bmatrix}3\\1\end{bmatrix}=\begin{bmatrix}3s\\s\end{bmatrix}\ ,\quad s\begin{bmatrix}2&2&7\\4&9&3\end{bmatrix}=\begin{bmatrix}2s&2s&7s\\4s&9s&3s\end{bmatrix}\ .$$

Multiplying two matrices or vectors is a somewhat more peculiar operation. Consider first multiplying two vectors, a row-vector and a column-vector, in this order. We get

$$\begin{bmatrix}a&b\end{bmatrix}\begin{bmatrix}c\\d\end{bmatrix}=ac+bd\ .$$

Generally, in the same way, we can multiply a $1\times k$ vector with a $k\times 1$ vector. Note that the "inner" dimensions (the k here) have to be the same for this to work. The result has the dimensions 1×1, i.e. it is a scalar. When multiplying a matrix with a matrix, we do the same operation just for each row of the first matrix and for each column of the second matrix. We get

$$\begin{bmatrix}a&b\\c&d\end{bmatrix}\begin{bmatrix}e&f&g\\h&i&j\end{bmatrix}=\begin{bmatrix}ae+bh&af+bi&ag+bj\\ce+dh&cf+di&cg+dj\end{bmatrix}\ .$$

In the same way, we can multiply a $n\times k$ matrix with a $k\times m$ matrix. The result is then $n\times m$. Again, the inner dimension (k here) have to correspond. Relatedly, note that the order by which two matrices are multiplied matters!

Matrix notation is useful due to its efficiency. For example, we can write

$$v_1b_1+v_2b_2+...+v_kb_k=\sum_{j=1}^{k}x_jb_j\quad\text{compactly as}\quad v'b\ ,$$

whereby $v=(v_1,v_2,...,v_k)'$ and $b=(b_1,b_2,...,b_k)'$ are both $k\times 1$ vectors. Similarly, we can write the three equations

$$x_i'b=0\ \text{ for }\ i=1,2,3\quad\text{compactly as}\quad Xb=0\ ,$$

whereby

$$X = \begin{bmatrix} x_1' \\ x_2' \\ x_3' \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} & ... & x_{1k} \\ x_{21} & x_{22} & ... & x_{2k} \\ x_{31} & x_{32} & ... & x_{3k} \end{bmatrix}$$

is a $3 \times k$ matrix that stacks the vectors x_1, x_2 and x_3 along rows, and, with slight abuse of notation, 0 is a 3×1 vector of zeros. This is an example how matrix notation allows us to write several (here three) equations as one. Also, note that

$$v'v = v_1^2 + v_2^2 + v_3^2$$

computes the sum of squares of the elements in vector v. As a result, we know that $v'v \geq 0$ for any vector v.

Above we multiplied a row-vector with a column-vector (of the same length) and obtained a scalar. When multiplying a column-vector with a row-vector instead, we obtain a matrix. (Note that we apply thereby the same rule as above, because this is nothing else than multiplying an $n \times 1$ matrix with a $1 \times m$ matrix.) For example:

$$\begin{bmatrix} a \\ b \end{bmatrix} \begin{bmatrix} c & d & e \end{bmatrix} = \begin{bmatrix} ac & ad & ae \\ bc & bd & be \end{bmatrix} .$$

Therefore, multiplying an $n \times 1$ vector with a $1 \times m$ vector, we get an $n \times m$ matrix. Using this result, we can see that

$$x_i x_i' = \begin{bmatrix} x_{i1} \\ x_{i2} \\ ... \\ x_{ik} \end{bmatrix} \begin{bmatrix} x_{i1} & x_{i2} & ... & x_{ik} \end{bmatrix} = \begin{bmatrix} x_{i1}^2 & x_{i1}x_{i2} & ... & x_{i1}x_{ik} \\ x_{i2}x_{i1} & x_{i2}^2 & ... & x_{i2}x_{ik} \\ \vdots & \vdots & \ddots & \vdots \\ x_{ik}x_{i1} & x_{ik}x_{i2} & ... & x_{ik}^2 \end{bmatrix}$$

gives a $k \times k$ matrix. We can then write

$$x_1x_1' + x_2x_2' + x_3x_3' = \sum_{i=1}^{3} x_i x_i' \quad \text{compactly as} \quad X'X ,$$

with X defined as above. Thereby,

$$X' = \begin{bmatrix} x_1 & x_2 & x_3 \end{bmatrix}$$

is a $k \times 3$ matrix that stacks the vectors x_1, x_2 and x_3 along columns.

Combining transposition and addition or multiplication, respectively, we have the rules

$$(AB)' = B'A' \quad \text{and} \quad (A+B)' = A' + B' \ .$$

A matrix M is said to be symmetric if $M' = M$, i.e. if transposing the matrix does not change it. Note that the matrix $X'X$ is symmetric. We say that a matrix M is square if it has the same number of rows and columns. Square matrices can be raised to powers; e.g. we can compute $M^2 = MM$ or $M^c = MM^{c-1}$. A square matrix D is diagonal if it has non-zero elements only along the diagonal. For example, we might have the $k \times k$ diagonal matrix

$$S = \begin{bmatrix} s_1 & 0 & ... & 0 \\ 0 & s_2 & ... & 0 \\ \vdots & \vdots & \ddots & \vdots \\ 0 & 0 & ... & s_k \end{bmatrix} ,$$

which we can also just write compactly as $S = diag(s_1, s_2, ..., s_k)$. A particularly important diagonal matrix is the identity matrix, $I = diag(1, 1, ..., 1)$ which has just ones on its diagonal. We sometimes specify its dimension using a subscript: I_k is a $k \times k$ identity matrix. It is easy to verify that for a square matrix S we have $S^c = diag(s_1^c, s_2^c, ..., s_k^c)$. In particular, $I^c = I$. A square matrix T is lower-triangular if it has non-zero elements only on and below its diagonal, and it is upper-triangular if it has non-zero elements only on and above its diagonal. For example, we could have

$$T_1 = \begin{bmatrix} 1 & 0 & 0 & 0 \\ 2 & 4 & 0 & 0 \\ -1 & 3 & 5 & 0 \\ 9 & -4 & 5 & 1 \end{bmatrix} , \quad T_2 = \begin{bmatrix} 1 & 4 & 1 & -4 \\ 0 & 4 & 5 & 0 \\ 0 & 0 & 5 & 6 \\ 0 & 0 & 0 & 3 \end{bmatrix} .$$

Clearly, transposing a lower-triangular matrix gives an upper-triangular matrix, and vice versa.

M3.2 Rank, Invertibility, Determinants and Eigenvalues

The vectors $v_1, v_2, ..., v_k$ are said to be linearly dependent if we can find scalars $a_1, a_2, ..., a_k$ – whereby at least one of them has to be non-zero – s.t.

$$a_1 v_1 + a_2 v_2 + ... + a_k v_k = 0 \ .$$

Otherwise they are linearly independent. Linear dependence implies that we can write at least one of the vectors (the one multiplied by a non-zero scalar) – w.l.o.g. the first vector, v_1 – as a linear combination of the others:

$$v_1 = \frac{-a_2}{a_1}v_2 + ... + \frac{-a_k}{a_1}v_k \ .$$

The rank of a matrix M is the number of linearly independent columns of M.[21] Loosely speaking, it tells us whether all the information encoded in M is important or we can reduce M to a smaller matrix (a matrix containing less vectors in its columns) without losing relevant information. It turns out that an $n \times m$ matrix, if $n \leq m$, can have at most rank n, and it can have at most rank m if $n \geq m$.[22] If a matrix (given its dimensions) has the largest possible rank, we say the matrix has full rank. Otherwise it is rank-deficient. It is cumbersome (and not necessary) to manually compute a matrix' rank. Instead, we can use the computer to compute it numerically, if needed.

Let M be a square $k \times k$ matrix. If it has full rank, then it is invertible, i.e. we can compute M^{-1}. This matrix is also $k \times k$, and we have $MM^{-1} = M^{-1}M = I$. Again, the manual procedures to compute an inverse are cumbersome,[23] but we can compute them easily numerically. However, it is important to know that $(sM)^{-1} = s^{-1}M^{-1}$. Also, it is worth knowing that for a diagonal matrix $S = diag(s_1, s_2, ..., s_k)$, we have $S^{-1} = diag(s_1^{-1}, s_2^{-1}, ..., s_k^{-1})$. In the same way, $I^{-1} = I$. Combining inversion and transition or multiplication, respectively, we have the rules

$$(A^{-1})' = (A')^{-1} \quad \text{and} \quad (AB)^{-1} = B^{-1}A^{-1} \ .$$

However, note that $(A + B)^{-1}$ cannot (in general) be simplified any further.

For a square matrix M, we can compute its determinant, denoted by $|M|$ or $det(M)$. It is a scalar, computed from the elements of M, and it is non-zero iff the matrix M has full rank (i.e. is invertible). Once again, it is cumbersome to compute manually, but easy to compute numerically. However, it is worth noting that for a diagonal matrix $S = diag(s_1, s_2, ..., s_k)$, the determinant is obtained by multiplying the diagonal entries: $|S| = \prod_{j=1}^{k} s_j$. Also, it is

[21] It is the space spanned by the columns of M.

[22] This is why the rank is the number of linearly independent columns as well as the number of linearly independent rows in M.

[23] In the case of a 2×2 matrix, however, we get

$$M^{-1} = \begin{bmatrix} a & b \\ c & d \end{bmatrix}^{-1} = \frac{1}{|M|}\begin{bmatrix} d & -b \\ -c & a \end{bmatrix} \ ,$$

where $|M| = ad - bc$ is the determinant of M. This can be useful for some theoretical insights, for which we can reduce our generally large matrices to 2×2 matrices.

useful to know that

$$|cM| = c^k|M|$$

for a scalar c and a $k \times k$ matrix M.

For a square matrix M, we can also compute its eigenvalues and corresponding eigenvectors (once again numerically). An eigenvalue λ and an eigenvector v of M satisfy the equation

$$Mv = \lambda v \ .^{24}$$

For a $k \times k$ matrix M, we can find k such eigenvalues $\lambda_1, \lambda_2, ..., \lambda_k$ with corresponding eigenvectors $v_1, v_2, ..., v_k$. It turns out that the determinant of M is the product of its eigenvalues: $|M| = \prod_{j=1}^{k} \lambda_j$. Eigenvalues are useful because they can be used to efficiently compute powers of a square matrix, and they tell us something about the properties of a matrix raised to higher and higher powers. It turns out that we can write

$$M = Q\Lambda Q^{-1} \ ,$$

where $\Lambda = diag(\lambda_1, \lambda_2, ..., \lambda_k)$ is a diagonal matrix containing the eigenvalues of M, and Q is a $k \times k$ matrix stacking the eigenvectors along columns. As a result, we have $M^c = Q\Lambda^c Q^{-1}$ for some integer c, where $\Lambda^c = diag(\lambda_1^c, \lambda_2^c, ..., \lambda_k^c)$. If all the eigenvalues of M are smaller than one in absolute value, then we know that $\lim_{h \to \infty} M^h = 0$.[25]

M3.3 Revisiting Functions

Consider a function that has multiple scalar-arguments and returns a single number (scalar). Equivalently, it can be thought of as a function that takes a single vector as its argument. For example, the function $f(x, y) = 2x + y$ can be written as $f(v) = \begin{bmatrix} 2 & 1 \end{bmatrix} v$, where $v = (x, y)'$ is a 2×1 vector that contains the scalars x and y. It maps $\mathbb{R}^2$ into $\mathbb{R}$. The function

$$f(b_1, b_2) = (y_1 - x_{11}b_1 - x_{12}b_2)^2 + (y_2 - x_{21}b_1 - x_{22}b_2)^2 + (y_3 - x_{31}b_1 - x_{32}b_2)^2$$

[24] Geometrically, the eigenvector v is a vector which does not change its direction when it is multiplied by the matrix M, but only gets scaled up or down by the eigenvalue λ.

[25] Note that the above results also imply that the eigenvalues of M^c are $\lambda_1^c, \lambda_2^c, ..., \lambda_k^c$. Also, we have $M^{-1} = Q^{-1}\Lambda^{-1}Q$, and the eigenvalues of M^{-1} are $\lambda_1^{-1}, \lambda_2^{-1}, ..., \lambda_k^{-1}$.

can be written as $f(b) = (y_1 - x_1'b)^2 + (y_2 - x_2'b)^2 + (y_3 - x_3'b)^2 = \sum_{i=1}^{3}(y_i - x_i'b)^2$, where $b = (b_1, b_2)'$. Even more compactly, this is

$$f(b) = (Y - Xb)'(Y - Xb) \ , \quad \text{where} \quad Y = \begin{bmatrix} y_1 \\ y_2 \\ y_3 \end{bmatrix} \ , \ X = \begin{bmatrix} x_1' \\ x_2' \\ x_3' \end{bmatrix} = \begin{bmatrix} x_{11} & x_{12} \\ x_{21} & x_{22} \\ x_{31} & x_{32} \end{bmatrix} \ .$$

This function maps $\mathbb{R}^2$ into $\mathbb{R}_+$.

Similarly, a function can also take a matrix as an argument. For example, we might have $f(M) = a'Mb$, i.e. we take the matrix M and multiply it (from the left) by the row-vector a' and from the left by the column-vector b. Note that the dimensions have to correspond. For example, if a is $n \times 1$ and b is $m \times 1$, then M has to be $n \times m$. The result is a scalar. The function maps $\mathbb{R}^n \times \mathbb{R}^m = \mathbb{R}^{nm}$ into $\mathbb{R}$.

In the same vein, a function can also have multiple outputs, i.e. return a vector or matrix as its output rather than a scalar. We speak of vector- and matrix-valued functions, respectively. For example, with a scalar-argument x, we might have

$$f(x) = \begin{bmatrix} x^2 \\ 2x \end{bmatrix} \ , \quad \text{or} \quad f(x) = \begin{bmatrix} x^2 & x^3 \\ 2x & 1 \end{bmatrix} \ .$$

Of course, we can also have a function that takes as input a matrix (or vector) and returns a matrix (or vector). For example, $f(v) = 2vv'$ takes as input a (column-)vector v and returns a matrix, or $f(M) = a'M$ takes as input a matrix M and multiplies it by a (row-)vector a' to return a row-vector. In its most general form, a (real-valued)[26] function maps a d-dimensional domain into a c-dimensional codomain, $f : \mathbb{R}^d \to \mathbb{R}^c$.

M3.4 Revisiting Differentiation

Consider a scalar-valued function that takes a vector as its argument (i.e. has two scalar-arguments). For example, take $f : \mathbb{R}^2 \to \mathbb{R}$ defined by $f(v_1, v_2) = 2v_1 + v_2$, which can be written as $f(v) = \begin{bmatrix} 2 & 1 \end{bmatrix} v$, where $v = (v_1, v_2)'$. We can obtain the partial derivatives of f w.r.t. all of its scalar-arguments in a single equation by taking the derivative of f w.r.t. its single vector-argument v:

$$\frac{\partial f}{\partial v} = \begin{bmatrix} \partial f / \partial v_1 \\ \partial f / \partial v_2 \end{bmatrix} = \begin{bmatrix} 2 \\ 1 \end{bmatrix} \ .$$

[26] See discussion above; we could in principle have functions that map any set into any other set. However, for our purposes it suffices to think of functions mapping real numbers into real numbers.

It is worth remembering that taking the derivative of a scalar-valued function f w.r.t. a (column-) vector v, we should get a (column-) vector of the same dimension. Each of its entries shows the derivative of f w.r.t. an individual element of v. Just as when taking derivatives w.r.t. a scalar, we can denote $\frac{\partial f}{\partial v}$ also as f' or $f^{(1)}$. The convention is to write $\frac{\partial f}{\partial v}$ rather than $\frac{\partial f}{\partial v'}$ because taking derivatives w.r.t. the column-vector v should give a column-vector (we go from top to bottom) and taking derivatives w.r.t. the row-vector v' should give a row-vector (we go from left to right). However, it is easy to see that transposing one we get the other.

Consider now taking derivatives of a vector-valued function. For illustration, let the function $g : \mathbb{R}^2 \to \mathbb{R}^3$ have a 2×1 argument (two scalar-arguments) and return a 3×1 output (three scalar-outputs). Specifically, let

$$g(v) = \begin{bmatrix} 2v_1 + v_2 \\ v_1^2 + \log v_2 \\ -v_2^3 \end{bmatrix}.$$

We can take the derivative of each of the three scalar-outputs of g w.r.t. each of the two scalar-arguments. Because $g(v)$ already returns a column-vector as its output, we have to take the derivatives w.r.t. v', i.e. go from left to right, rather than w.r.t. v, because we cannot go from top to bottom. We get

$$\frac{\partial g}{\partial v'} = \begin{bmatrix} \partial g_1/\partial v_1 & \partial g_1/\partial v_2 \\ \partial g_2/\partial v_1 & \partial g_2/\partial v_2 \\ \partial g_3/\partial v_1 & \partial g_3/\partial v_2 \end{bmatrix} = \begin{bmatrix} 2 & 1 \\ 2v_1 & v_2^{-1} \\ 0 & -3v_2^2 \end{bmatrix}.$$

Thereby, g_1, g_2 and g_3 denote the three elements of the output of g.

This discussion suggests that we can write the matrix of second-order derivatives of a scalar-valued function $f : \mathbb{R}^k \to \mathbb{R}$ as the $k \times k$ matrix

$$\frac{\partial^2 f}{\partial v \partial v'} = \frac{\partial}{\partial v'}\frac{\partial f}{\partial v} = \begin{bmatrix} \frac{\partial^2 f}{\partial v_1 \partial v_1} & \frac{\partial^2 f}{\partial v_1 \partial v_2} & \cdots & \frac{\partial^2 f}{\partial v_1 \partial v_k} \\ \frac{\partial^2 f}{\partial v_2 \partial v_1} & \frac{\partial^2 f}{\partial v_2 \partial v_2} & \cdots & \frac{\partial^2 f}{\partial v_2 \partial v_k} \\ \vdots & \vdots & \ddots & \vdots \\ \frac{\partial^2 f}{\partial v_k \partial v_1} & \frac{\partial^2 f}{\partial v_k \partial v_2} & \cdots & \frac{\partial^2 f}{\partial v_k \partial v_k} \end{bmatrix},$$

i.e. we first compute the derivative of f w.r.t. its $k \times 1$ vector of arguments v – which gives us a $k \times 1$ vector of first derivatives – and we then compute the derivative of each of its elements w.r.t. the $1 \times k$ vector v' – which gives us overall a $k \times k$ matrix of second derivatives. Again, we can write $\frac{\partial^2 f}{\partial v \partial v'}$ also as f'' or $f^{(2)}$. As an example, consider the function

$f(v_1, v_2) = v_1^2 + \log v_2 - v_1 v_2^2$. We get

$$f' = \begin{bmatrix} 2v_1 - v_2^2 \\ v_2^{-1} - 2v_1 v_2 \end{bmatrix}, \quad \text{and} \quad f'' = \begin{bmatrix} 2 & -2v_2 \\ -2v_2 & -v_2^{-2} - 2v_1 \end{bmatrix}.$$

Note that the matrix of second-order derivatives is symmetric because $\frac{\partial f}{\partial v_i \partial v_j} = \frac{\partial f}{\partial v_j \partial v_i}$. There is no compact way to write third-order derivatives of a scalar-valued function in matrix-notation (we would get an array, i.e. a three-dimensional object). However, in econometrics we rarely need to go beyond the first two orders.

For a scalar-valued function that takes a matrix as its argument, we can arrange its first-order derivatives as a matrix. As an example, consider $f : \mathbb{R}^4 \to \mathbb{R}$ defined by $f(M) = a'Mb$ (i.e. M is a 2×2 matrix so that the domain of f is four-dimensional). Note that written out, we have

$$\begin{aligned} f(M) &= \begin{bmatrix} a_1 & a_2 \end{bmatrix} \begin{bmatrix} M_{11} & M_{12} \\ M_{21} & M_{22} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \\ &= \begin{bmatrix} a_1 M_{11} + a_2 M_{21} & a_1 M_{12} + a_2 M_{22} \end{bmatrix} \begin{bmatrix} b_1 \\ b_2 \end{bmatrix} \\ &= b_1 a_1 M_{11} + b_1 a_2 M_{21} + b_2 a_1 M_{12} + b_2 a_2 M_{22} . \end{aligned}$$

For the first-order derivatives, we then get

$$f' = \frac{\partial f}{\partial M} = \begin{bmatrix} \partial f / \partial M_{11} & \partial f / \partial M_{12} \\ \partial f / \partial M_{21} & \partial f / \partial M_{22} \end{bmatrix} = \begin{bmatrix} b_1 a_1 & b_2 a_1 \\ b_2 a_2 & b_2 a_3 \end{bmatrix}.$$

It turns out that this is simply $f' = ab'$.

Just like when taking derivatives w.r.t. a scalar, there are many results that state what the derivative of a given function w.r.t. a vector or matrix is. Above we saw that the derivative of $f(M) = a'Mb$ is $f' = ab'$. Similarly, for $f(v) = v'Mv$ we have $f' = (M + M')v$, or for $f(v) = (v - s)'M(v - s)$ we have $f' = 2M(v - s)$ and $f'' = 2M'$. These and many other useful results can be looked up e.g. in Petersen and Pedersen (2012).[27]

[27] While some of these results are easy to see – e.g. $f(v) = a'v$ leads to $f' = a$ – most are not. However, they are easily verified, and in practice one simply uses these results without re-deriving them anew for every problem, just as for derivatives w.r.t. a scalar, like $f(x) = \log x$ leading to $f' = x^{-1}$.