Chapter 2
Notation and Norms

2.1 Introduction

This chapter recalls the usual convention for distinguishing scalars, vectors, and matrices. Vetter's notation for matrix derivatives is then explained, as well as the meaning of the expressions *little o* and *big O* employed for comparing the local or asymptotic behaviors of functions. The most important vector and matrix norms are finally described. Norms find a first application in the definition of types of convergence speeds for iterative algorithms.

2.2 Scalars, Vectors, and Matrices

Unless stated otherwise, scalar variables are real valued, as are the entries of vectors and matrices.

Italics are for *scalar variables* (v or V), bold lower-case letters for *column vectors* ($\mathbf{v}$), and bold upper-case letters for *matrices* ($\mathbf{M}$). *Transposition*, the transformation of columns into rows in a vector or matrix, is denoted by the superscript $^{\mathrm{T}}$. It applies to what is to its *left*, so $\mathbf{v}^{\mathrm{T}}$ is a *row vector* and, in $\mathbf{A}^{\mathrm{T}}\mathbf{B}$, $\mathbf{A}$ is transposed, not $\mathbf{B}$.

The *identity matrix* is $\mathbf{I}$, with $\mathbf{I}_n$ the $(n \times n)$ identity matrix. The ith column vector of $\mathbf{I}$ is the *canonical vector* $\mathbf{e}^i$.

The entry at the intersection of the ith row and jth column of $\mathbf{M}$ is $m_{i,j}$. The product of matrices

$$\mathbf{C} = \mathbf{AB} \tag{2.1}$$

thus implies that

$$c_{i,j} = \sum_k a_{i,k} b_{k,j}, \tag{2.2}$$

É. Walter, *Numerical Methods and Optimization*,
DOI: 10.1007/978-3-319-07671-3_2,

and the number of columns in $\mathbf{A}$ must be equal to the number of rows in $\mathbf{B}$. Recall that the product of matrices (or vectors) is *not commutative*, in general. Thus, for instance, when $\mathbf{v}$ and $\mathbf{w}$ are columns vectors with the same dimension, $\mathbf{v}^{\mathrm{T}}\mathbf{w}$ is a scalar whereas $\mathbf{w}\mathbf{v}^{\mathrm{T}}$ is a (rank-one) square matrix.

Useful relations are

$$(\mathbf{AB})^{\mathrm{T}} = \mathbf{B}^{\mathrm{T}}\mathbf{A}^{\mathrm{T}}, \tag{2.3}$$

and, provided that $\mathbf{A}$ and $\mathbf{B}$ are invertible,

$$(\mathbf{AB})^{-1} = \mathbf{B}^{-1}\mathbf{A}^{-1}. \tag{2.4}$$

If $\mathbf{M}$ is square and symmetric, then all of its eigenvalues are real. $\mathbf{M} \succ 0$ then means that each of these eigenvalues is strictly positive ($\mathbf{M}$ is *positive definite*), while $\mathbf{M} \succcurlyeq 0$ allows some of them to be zero ($\mathbf{M}$ is *non-negative definite*).

2.3 Derivatives

Provided that $f(\cdot)$ is a sufficiently differentiable function from $\mathbb{R}$ to $\mathbb{R}$,

$$\dot{f}(x) = \frac{\mathrm{d}f}{\mathrm{d}x}(x), \tag{2.5}$$

$$\ddot{f}(x) = \frac{\mathrm{d}^2 f}{\mathrm{d}x^2}(x), \tag{2.6}$$

$$f^{(k)}(x) = \frac{\mathrm{d}^k f}{\mathrm{d}x^k}(x). \tag{2.7}$$

Vetter's notation [1] will be used for derivatives of matrices with respect to matrices. (A word of caution is in order: there are other, incompatible notations, and one should be cautious about mixing formulas from different sources.)

If $\mathbf{A}$ is $(n_{\mathrm{A}} \times m_{\mathrm{A}})$ and $\mathbf{B}$ $(n_{\mathrm{B}} \times m_{\mathrm{B}})$, then

$$\mathbf{M} = \frac{\partial \mathbf{A}}{\partial \mathbf{B}} \tag{2.8}$$

is an $(n_{\mathrm{A}} n_{\mathrm{B}} \times m_{\mathrm{A}} m_{\mathrm{B}})$ matrix, such that the $(n_{\mathrm{A}} \times m_{\mathrm{A}})$ submatrix in position (i, j) is

$$\mathbf{M}_{i,j} = \frac{\partial \mathbf{A}}{\partial b_{i,j}}. \tag{2.9}$$

Remark 2.1 $\mathbf{A}$ and $\mathbf{B}$ in (2.8) may be row or column vectors. □

Example 2.1 If $\mathbf{v}$ is a generic column vector of $\mathbb{R}^n$, then

$$\frac{\partial \mathbf{v}}{\partial \mathbf{v}^{\mathrm{T}}} = \frac{\partial \mathbf{v}^{\mathrm{T}}}{\partial \mathbf{v}} = \mathbf{I}_n. \tag{2.10}$$

□

Example 2.2 If $J(\cdot)$ is a differentiable function from $\mathbb{R}^n$ to $\mathbb{R}$, and $\mathbf{x}$ a vector of $\mathbb{R}^n$, then

$$\frac{\partial J}{\partial \mathbf{x}}(\mathbf{x}) = \begin{bmatrix} \frac{\partial J}{\partial x_1} \\ \frac{\partial J}{\partial x_2} \\ \vdots \\ \frac{\partial J}{\partial x_n} \end{bmatrix} (\mathbf{x}) \tag{2.11}$$

is a column vector, called the *gradient* of $J(\cdot)$ at $\mathbf{x}$. □

Example 2.3 If $J(\cdot)$ is a twice differentiable function from $\mathbb{R}^n$ to $\mathbb{R}$, and $\mathbf{x}$ a vector of $\mathbb{R}^n$, then

$$\frac{\partial^2 J}{\partial \mathbf{x} \partial \mathbf{x}^{\mathrm{T}}}(\mathbf{x}) = \begin{bmatrix} \frac{\partial^2 J}{\partial x_1^2} & \frac{\partial^2 J}{\partial x_1 \partial x_2} & \cdots & \frac{\partial^2 J}{\partial x_1 \partial x_n} \\ \frac{\partial^2 J}{\partial x_2 \partial x_1} & \frac{\partial^2 J}{\partial x_2^2} & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{\partial^2 J}{\partial x_n \partial x_1} & \cdots & \cdots & \frac{\partial^2 J}{\partial x_n^2} \end{bmatrix} (\mathbf{x}) \tag{2.12}$$

is an $(n \times n)$ matrix, called the *Hessian* of $J(\cdot)$ at $\mathbf{x}$. Schwarz's theorem ensures that

$$\frac{\partial^2 J}{\partial x_i \partial x_j}(\mathbf{x}) = \frac{\partial^2 J}{\partial x_j \partial x_i}(\mathbf{x}), \tag{2.13}$$

provided that both are continuous at $\mathbf{x}$ and $\mathbf{x}$ belongs to an open set in which both are defined. Hessians are thus symmetric, except in pathological cases not considered here. □

Example 2.4 If $\mathbf{f}(\cdot)$ is a differentiable function from $\mathbb{R}^n$ to $\mathbb{R}^p$, and $\mathbf{x}$ a vector of $\mathbb{R}^n$, then

$$\mathbf{J}(\mathbf{x}) = \frac{\partial \mathbf{f}}{\partial \mathbf{x}^{\mathrm{T}}}(\mathbf{x}) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & \frac{\partial f_2}{\partial x_2} & & \vdots \\ \vdots & & \ddots & \vdots \\ \frac{\partial f_p}{\partial x_1} & \cdots & \cdots & \frac{\partial f_p}{\partial x_n} \end{bmatrix} \tag{2.14}$$

is the $(p \times n)$ *Jacobian matrix* of $\mathbf{f}(\cdot)$ at $\mathbf{x}$. When $p = n$, the Jacobian matrix is square and its determinant is the *Jacobian*. □

Remark 2.2 The last three examples show that the Hessian of $J(\cdot)$ at $\mathbf{x}$ is the Jacobian matrix of its gradient function evaluated at $\mathbf{x}$. □

Remark 2.3 Gradients and Hessians are frequently used in the context of optimization, and Jacobian matrices when solving systems of nonlinear equations. □

Remark 2.4 The *Nabla operator* ∇, a vector of partial derivatives with respect to all the variables of the function on which it operates

$$\nabla = \left(\frac{\partial}{\partial x_1}, \ldots, \frac{\partial}{\partial x_n} \right)^{\mathrm{T}}, \tag{2.15}$$

is often used to make notation more concise, especially for partial differential equations. Applying ∇ to a scalar function J and evaluating the result at $\mathbf{x}$, one gets the gradient vector

$$\nabla J(\mathbf{x}) = \frac{\partial J}{\partial \mathbf{x}}(\mathbf{x}). \tag{2.16}$$

If the scalar function is replaced by a vector function $\mathbf{f}$, one gets the Jacobian matrix

$$\nabla \mathbf{f}(\mathbf{x}) = \frac{\partial \mathbf{f}}{\partial \mathbf{x}^{\mathrm{T}}}(\mathbf{x}), \tag{2.17}$$

where $\nabla \mathbf{f}$ is interpreted as $\left(\nabla \mathbf{f}^{\mathrm{T}}\right)^{\mathrm{T}}$.

By applying ∇ twice to a scalar function J and evaluating the result at $\mathbf{x}$, one gets the Hessian matrix

$$\nabla^2 J(\mathbf{x}) = \frac{\partial^2 J}{\partial \mathbf{x} \partial \mathbf{x}^{\mathrm{T}}}(\mathbf{x}). \tag{2.18}$$

(∇^2 is sometimes taken to mean the *Laplacian operator* Δ, such that

$$\Delta f(\mathbf{x}) = \sum_{i=1}^{n} \frac{\partial^2 f}{\partial x_i^2}(\mathbf{x}) \tag{2.19}$$

is a scalar. The context and dimensional considerations should make what is meant clear.) □

Example 2.5 If $\mathbf{v}$, $\mathbf{M}$, and $\mathbf{Q}$ do not depend on $\mathbf{x}$ and $\mathbf{Q}$ is symmetric, then

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{v}^{\mathrm{T}}\mathbf{x}) = \mathbf{v}, \tag{2.20}$$

$$\frac{\partial}{\partial \mathbf{x}^{\mathrm{T}}}(\mathbf{M}\mathbf{x}) = \mathbf{M}, \tag{2.21}$$

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^{\mathrm{T}}\mathbf{M}\mathbf{x}) = (\mathbf{M} + \mathbf{M}^{\mathrm{T}})\mathbf{x} \tag{2.22}$$

and

$$\frac{\partial}{\partial \mathbf{x}}(\mathbf{x}^{\mathrm{T}}\mathbf{Q}\mathbf{x}) = 2\mathbf{Q}\mathbf{x}. \tag{2.23}$$

These formulas will be used quite frequently. □

2.4 Little o and Big O

The function $f(x)$ is $\mathrm{o}(g(x))$ as x tends to x_0 if

$$\lim_{x \to x_0} \frac{f(x)}{g(x)} = 0, \tag{2.24}$$

so $f(x)$ gets negligible compared to $g(x)$ for x sufficiently close to x_0. In what follows, x_0 is always taken equal to zero, so this need not be specified, and we just write $f(x) = \mathrm{o}(g(x))$.

The function $f(x)$ is $\mathrm{O}(g(x))$ as x tends to *infinity* if there exists real numbers x_0 and M such that

$$x > x_0 \Rightarrow |f(x)| \leqslant M|g(x)|. \tag{2.25}$$

The function $f(x)$ is $\mathrm{O}(g(x))$ as x tends to *zero* if there exists real numbers δ and M such that

$$|x| < \delta \Rightarrow |f(x)| \leqslant M|g(x)|. \tag{2.26}$$

The notation $\mathrm{O}(x)$ or $\mathrm{O}(n)$ will be used in two contexts:

- when dealing with Taylor expansions, x is a real number tending to zero,
- when analyzing algorithmic complexity, n is a positive integer tending to infinity.

Example 2.6 The function

$$f(x) = \sum_{i=2}^{m} a_i x^i,$$

with $m \geqslant 2$, is such that

$$\lim_{x \to 0} \frac{f(x)}{x} = \lim_{x \to 0} \left(\sum_{i=2}^{m} a_i x^{i-1} \right) = 0,$$

so $f(x) = \mathrm{o}(x)$ when x tends to zero. Now, if $|x| < 1$, then

$$\frac{|f(x)|}{x^2} < \sum_{i=2}^{m} |a_i|,$$

so $f(x) = \mathrm{O}(x^2)$ when x tends to zero. If, on the other hand, x is taken equal to the (large) positive integer n, then

$$f(n) = \sum_{i=2}^{m} a_i n^i \leqslant \sum_{i=2}^{m} |a_i n^i| \leqslant \left(\sum_{i=2}^{m} |a_i| \right) \cdot n^m,$$

so $f(n) = \mathrm{O}(n^m)$ when n tends to infinity. □

2.5 Norms

A function $f(\cdot)$ from a vector space $\mathbb{V}$ to $\mathbb{R}$ is a *norm* if it satisfies the following three properties:

1. $f(\mathbf{v}) \geqslant 0$ for all $\mathbf{v} \in \mathbb{V}$ (*positivity*),
2. $f(\alpha \mathbf{v}) = |\alpha| \cdot f(\mathbf{v})$ for all $\alpha \in \mathbb{R}$ and $\mathbf{v} \in \mathbb{V}$ (*positive scalability*),
3. $f(\mathbf{v}^1 \pm \mathbf{v}^2) \leqslant f(\mathbf{v}^1) + f(\mathbf{v}^2)$ for all $\mathbf{v}^1 \in \mathbb{V}$ and $\mathbf{v}^2 \in \mathbb{V}$ (*triangle inequality*).

These properties imply that $f(\mathbf{v}) = 0 \Rightarrow \mathbf{v} = \mathbf{0}$ (*non-degeneracy*). Another useful relation is

$$|f(\mathbf{v}^1) - f(\mathbf{v}^2)| \leqslant f(\mathbf{v}^1 \pm \mathbf{v}^2). \tag{2.27}$$

Norms are used to quantify distances between vectors. They play an essential role, for instance, in the characterization of the intrinsic difficulty of numerical problems via the notion of condition number (see Sect. 3.3) or in the definition of cost functions for optimization.

2.5.1 Vector Norms

The most commonly used norms in $\mathbb{R}^n$ are the l_p norms

$$\|\mathbf{v}\|_p = \left(\sum_{i=1}^{n} |v_i|^p \right)^{\frac{1}{p}}, \tag{2.28}$$

with $p \geqslant 1$. They include

- the *Euclidean norm* (or l_2 norm)

$$||\mathbf{v}||_2 = \sqrt{\sum_{i=1}^{n} v_i^2} = \sqrt{\mathbf{v}^{\mathrm{T}}\mathbf{v}}, \tag{2.29}$$

- the *taxicab norm* (or Manhattan norm, or grid norm, or l_1 norm)

$$||\mathbf{v}||_1 = \sum_{i=1}^{n} |v_i|, \tag{2.30}$$

- the *maximum norm* (or l_∞ norm, or Chebyshev norm, or uniform norm)

$$||\mathbf{v}||_\infty = \max_{1 \leqslant i \leqslant n} |v_i|. \tag{2.31}$$

They are such that

$$||\mathbf{v}||_2 \leqslant ||\mathbf{v}||_1 \leqslant n||\mathbf{v}||_\infty, \tag{2.32}$$

and

$$\mathbf{v}^{\mathrm{T}}\mathbf{w} \leqslant \|\mathbf{v}\|_2 \cdot \|\mathbf{w}\|_2. \tag{2.33}$$

The latter result is known as the *Cauchy-Schwarz inequality*.

Remark 2.5 If the entries of $\mathbf{v}$ were complex, norms would be defined differently. The Euclidean norm, for instance, would become

$$||\mathbf{v}||_2 = \sqrt{\mathbf{v}^{\mathrm{H}}\mathbf{v}}, \tag{2.34}$$

where $\mathbf{v}^{\mathrm{H}}$ is the *transconjugate* of $\mathbf{v}$, i.e., the row vector obtained by transposing the column vector $\mathbf{v}$ and replacing each of its entries by its complex conjugate. □

Example 2.7 For the complex vector

$$\mathbf{v} = \begin{bmatrix} a \\ a\mathrm{i} \end{bmatrix},$$

where a is some nonzero real number and i is the imaginary unit (such that $\mathrm{i}^2 = -1$), $\mathbf{v}^{\mathrm{T}}\mathbf{v} = 0$. This proves that $\sqrt{\mathbf{v}^{\mathrm{T}}\mathbf{v}}$ is not a norm. The value of the Euclidean norm of $\mathbf{v}$ is $\sqrt{\mathbf{v}^{\mathrm{H}}\mathbf{v}} = \sqrt{2}|a|$. □

Remark 2.6 The so-called l_0 norm of a vector is the number of its nonzero entries. Used in the context of sparse estimation, where one is looking for an estimated parameter vector with as few nonzero entries as possible, it is *not* a norm, as it does not satisfy the property of positive scalability. □

2.5.2 Matrix Norms

Each vector norm *induces* a *matrix norm*, defined as

$$||\mathbf{M}|| = \max_{||\mathbf{v}||=1} ||\mathbf{M}\mathbf{v}||, \tag{2.35}$$

so

$$\|\mathbf{M}\mathbf{v}\| \leqslant \|\mathbf{M}\| \cdot \|\mathbf{v}\| \tag{2.36}$$

for any $\mathbf{M}$ and $\mathbf{v}$ for which the product $\mathbf{M}\mathbf{v}$ makes sense. This matrix norm is *subordinate* to the vector norm inducing it. The matrix and vector norms are then said to be *compatible*, an important property for the study of products of matrices and vectors.

- The matrix norm induced by the vector norm l_2 is the *spectral norm*, or *2-norm* ,

$$||\mathbf{M}||_2 = \sqrt{\rho(\mathbf{M}^{\mathrm{T}}\mathbf{M})}, \tag{2.37}$$

  where $\rho(\cdot)$ is the function that computes the *spectral radius* of its argument, i.e., the modulus of the eigenvalue(s) with the largest modulus. Since all the eigenvalues of $\mathbf{M}^{\mathrm{T}}\mathbf{M}$ are real and non-negative, $\rho(\mathbf{M}^{\mathrm{T}}\mathbf{M})$ is the largest of these eigenvalues. Its square root is the largest *singular value* of $\mathbf{M}$, denoted by $\sigma_{\max}(\mathbf{M})$. So

$$||\mathbf{M}||_2 = \sigma_{\max}(\mathbf{M}). \tag{2.38}$$

- The matrix norm induced by the vector norm l_1 is the *1-norm*

$$||\mathbf{M}||_1 = \max_j \sum_i |m_{i,j}|, \tag{2.39}$$

  which amounts to summing the absolute values of the entries of each *column* in turn and keeping the largest result.
- The matrix norm induced by the vector norm l_∞ is the *infinity norm*

$$||\mathbf{M}||_\infty = \max_i \sum_j |m_{i,j}|, \tag{2.40}$$

  which amounts to summing the absolute values of the entries of each *row* in turn and keeping the largest result. Thus

$$||\mathbf{M}||_1 = ||\mathbf{M}^{\mathrm{T}}||_\infty. \tag{2.41}$$

Since each subordinate matrix norm is compatible with its inducing vector norm,

$$||\mathbf{v}||_1 \text{ is compatible with } ||\mathbf{M}||_1, \tag{2.42}$$
$$||\mathbf{v}||_2 \text{ is compatible with } ||\mathbf{M}||_2, \tag{2.43}$$
$$||\mathbf{v}||_\infty \text{ is compatible with } ||\mathbf{M}||_\infty. \tag{2.44}$$

The *Frobenius norm*

$$||\mathbf{M}||_\mathrm{F} = \sqrt{\sum_{i,j} m_{i,j}^2} = \sqrt{\text{trace}\left(\mathbf{M}^\mathrm{T}\mathbf{M}\right)} \tag{2.45}$$

deserves a special mention, as it is not induced by any vector norm yet

$$||\mathbf{v}||_2 \text{ is compatible with } ||\mathbf{M}||_\mathrm{F}. \tag{2.46}$$

Remark 2.7 To evaluate a vector or matrix norm with MATLAB (or any other interpreted language based on matrices), it is much more efficient to use the corresponding dedicated function than to access the entries of the vector or matrix individually to implement the norm definition. Thus, `norm(X,p)` returns the p-norm of `X`, which may be a vector or a matrix, while `norm(M,'fro')` returns the Frobenius norm of the matrix `M`. □

2.5.3 Convergence Speeds

Norms can be used to study how quickly an iterative method would converge to the solution $\mathbf{x}^\star$ if computation were exact. Define the error at iteration k as

$$\mathbf{e}^k = \mathbf{x}^k - \mathbf{x}^\star, \tag{2.47}$$

where $\mathbf{x}^k$ is the estimate of $\mathbf{x}^\star$ at iteration k. The *asymptotic* convergence speed is *linear* if

$$\lim_{k\to\infty} \sup \frac{\|\mathbf{e}^{k+1}\|}{\|\mathbf{e}^k\|} = \alpha < 1, \tag{2.48}$$

with α the rate of convergence.

It is *superlinear* if

$$\lim_{k\to\infty} \sup \frac{\|\mathbf{e}^{k+1}\|}{\|\mathbf{e}^k\|} = 0, \tag{2.49}$$

and *quadratic* if

$$\lim_{k\to\infty} \sup \frac{\|\mathbf{e}^{k+1}\|}{\|\mathbf{e}^k\|^2} = \alpha < \infty. \tag{2.50}$$

A method with quadratic convergence thus also has superlinear and linear convergence. It is customary, however, to qualify a method with the best convergence it achieves. Quadratic convergence is better that superlinear convergence, which is better than linear convergence.

Remember that these convergence speeds are asymptotic, valid when the error has become small enough, and that they do not take the effect of rounding into account. They are meaningless if the initial vector $\mathbf{x}^0$ was too badly chosen for the method to converge to $\mathbf{x}^\star$. When the method does converge to $\mathbf{x}^\star$, they may not describe accurately its initial behavior and will no longer be true when rounding errors become predominant. They are nevertheless an interesting indication of what can be expected at best.

Reference

1. Vetter, W.: Derivative operations on matrices. IEEE Trans. Autom. Control **15**, 241–244 (1970)

http://www.springer.com/978-3-319-07670-6

Numerical Methods and Optimization
A Consumer Guide
Walter, É.
2014, XV, 476 p. 67 illus., 23 illus. in color., Hardcover
ISBN: 978-3-319-07670-6