Chapter 3
Mathematical Writing

Suppose you want to teach the "cat" concept to a very young child.
Do you explain that a cat is a relatively small,
primarily carnivorous mammal with retractile claws,
a distinctive sonic output, etc.?
I'll bet not.
You probably show the kid a lot of different cats,
saying "kitty" each time, until it gets the idea.
To put it more generally,
generalizations are best made by abstraction from experience.

— RALPH P. BOAS, *Can We Make Mathematics Intelligible?* (1981)

A good notation should be unambiguous, pregnant, easy to remember;
it should avoid harmful second meanings, and
take advantage of useful second meanings;
the order and connection of signs should
suggest the order and connection of things.

— GEORGE POLYA, *How to Solve It* (1957)

We have not succeeded in finding or constructing
a definition which starts out
"A Bravais lattice is . . . ";
the sources we have looked at say
"That was a Bravais lattice."

— CHARLES KITTEL, *Introduction to Solid State Physics* (1971)

Notation is everything.

— CHARLES F. VAN LOAN, *FFTs and the Sparse Factorization Idea* (1992)

The mathematical writer needs to be aware of a number of matters specific to mathematical writing, ranging from general issues, such as choice of notation, to particular details, such as how to punctuate mathematical expressions. In this chapter I begin by discussing some of the general issues and then move on to specifics.

3.1. What Is a Theorem?

What are the differences between theorems, lemmas, and propositions? To some extent, the answer depends on the context in which a result appears. Generally, a *theorem* is a major result that is of independent interest. The proof of a theorem is usually nontrivial. A *lemma*[3] is an auxiliary result—a stepping stone towards a theorem. Its proof may be easy or difficult. A straightforward and independent result that is worth encapsulating but that does not merit the title of a theorem may also be called a lemma. Indeed, there are some famous lemmas, such as the Riemann–Lebesgue Lemma in the theory of Fourier series and Farkas's Lemma in the theory of constrained optimization. Whether a result should be stated formally as a lemma or simply mentioned in the text depends on the level at which you are writing. In a research paper in linear algebra it would be inappropriate to give a lemma stating that the eigenvalues of a symmetric positive definite matrix are positive, as this standard result is so well known; but in a textbook for undergraduates it would be sensible to formalize this result.

It is not advisable to label all your results theorems, because if you do so you miss the opportunity to emphasize the logical structure of your work and to direct attention to the most important results. If you are in doubt about whether to call a result a lemma or a theorem, call it a lemma.

The term *proposition* is less widely used than lemma and theorem and its meaning is less clear. It tends to be used as a way to denote a minor theorem. Lecturers and textbook authors might feel that the modest tone of its name makes a proposition appear less daunting to students than a theorem. However, a proposition is not, as one student thought, "a theorem that might not be true".

A *corollary* is a direct or easy consequence of a lemma, theorem or proposition. It is important to distinguish between a corollary, which does not imply the parent result from which it came, and an extension or generalization of a result. Be careful not to over-glorify a corollary by failing to label it as such, for this gives it false prominence and obscures the role of the parent result.

[3]The plural of lemma is lemmata, or, more commonly, lemmas.

How many results are formally stated as lemmas, theorems, propositions or corollaries is a matter of personal style. Some authors develop their ideas in a sequence of results and proofs interspersed with definitions and comments. At the other extreme, some authors state very few results formally. A good example of the latter style is the classic book *The Algebraic Eigenvalue Problem* [296] by Wilkinson, in which only four titled theorems are given in 662 pages. As Boas [33] notes, "A great deal can be accomplished with arguments that fall short of being formal proofs."

A fifth kind of statement used in mathematical writing is a *conjecture*—a statement that the author thinks may be true but has been unable to prove or disprove. The author will usually have some strong evidence for the veracity of the statement. A famous example of a conjecture is the Goldbach conjecture (1742), which states that every even number greater than 2 is the sum of two primes; this is still unproved. One computer scientist (let us call him Alpha) joked in a talk "This is the Alpha and Beta conjecture. If it turns out to be false I would like it to be known as Beta's conjecture." However, it is not necessarily a bad thing to make a conjecture that is later disproved: identifying the question that the conjecture aims to answer can be an important contribution.

A *hypothesis* is a statement that is taken as a basis for further reasoning, usually in a proof—for example, an induction hypothesis. Hypotheses that stand on their own are uncommon; two examples are the Riemann hypothesis and the continuum hypothesis.

3.2. Proofs

Readers are often not very interested in the details of a proof but want to know the outline and the key ideas. They hope to learn a technique or principle that can be applied in other situations. When readers do want to study the proof in detail they naturally want to understand it with the minimum of effort. To help readers in both circumstances, it is important to emphasize the structure of a proof, the ease or difficulty of each step, and the key ideas that make it work. Here are some examples of the sorts of phrases that can be used (most of these are culled from proofs by Parlett in [217]).

> The aim/idea is to
> Our first goal is to show that
> Now for the harder part.
> The trick of the proof is to find
> ... is the key relation.
> The only, but crucial use of ... is that

To obtain ... a little manipulation is needed.
The essential observation is that

When you omit part of a proof it is best to indicate the nature and length of the omission, via phrases such as the following.

It is easy/simple/straightforward to show that
Some tedious manipulation yields
An easy/obvious induction gives
After two applications of ... we find
An argument similar to the one used in ... shows that

You should also strive to keep the reader informed of where you are in the proof and what remains to be done. Useful phrases include

First, we establish that
Our task is now to
Our problem reduces to
It remains to show that
We are almost ready to invoke
We are now in a position to
Finally, we have to show that

The end of a proof is often marked by the halmos symbol □ (see the quote on page 24). Sometimes the abbreviation QED (Latin: quod erat demonstrandum = which was to be demonstrated) is used instead.

There is much more to be said about writing (and devising) proofs. References include Franklin and Daoud [85], Garnier and Taylor [101], Lamport [173], Leron [177] and Polya [228].

3.3. The Role of Examples

A pedagogical tactic that is applicable to all forms of technical writing (from teaching to research) is to discuss specific examples before the general case. It is tempting, particularly for mathematicians, to adopt the opposite approach, but beginning with examples is often the more effective way to explain (see Boas's article [33] and the quote from it at the beginning of this chapter, a quote that itself illustrates this principle!).

A good example of how to begin with a specific case is provided by Strang in Chapter 1 of *Introduction to Applied Mathematics* [262]:

The simplest model in applied mathematics is a system of linear equations. It is also by far the most important, and we begin

this book with an extremely modest example:

$$2x_1 + 4x_2 = 2,$$
$$4x_1 + 11x_2 = 1.$$

After some further introductory remarks, Strang goes on to study in detail both this 2×2 system and a particular 4×4 system. General $n \times n$ matrices appear only several pages later.

Another example is provided by Watkins's *Fundamentals of Matrix Computations* [289]. Whereas most linear algebra textbooks introduce Gaussian elimination for general matrices before discussing Cholesky factorization for symmetric positive definite matrices, Watkins reverses the order, giving the more specific but algorithmically more straightforward method first.

An exercise in a textbook is a form of example. I saw a telling criticism in one book review that complained "The first exercise in the book was pointless, so why do the others?" To avoid such criticism, it is important to choose exercises and examples that have a clear purpose and illustrate a point. The first few exercises and examples should be among the best, to gain the reader's confidence. The same reviewer complained of another book that "it hides information in exercises and contains exercises that are too difficult." Whether such criticism is valid depends on your opinion of what are the key issues to be transmitted to the reader and on the level of the readership. Again, it helps to bear such potential criticism in mind when you write.

3.4. Definitions

Three questions to be considered when formulating a definition are "why?", "where?" and "how?" First, ask yourself why you are making a definition: is it really necessary? Inappropriate definitions can complicate a presentation and too many can overwhelm a reader, so it is wise to imagine yourself being charged a large sum for each one. Instead of defining a square matrix A to be contractive with respect to a norm $\|\cdot\|$ if $\|A\| < 1$, which is not a standard definition, you could simply say "A with $\|A\| < 1$" whenever necessary. This is easy to do if the property is needed on only a few occasions, and saves the reader having to remember what "contractive" means. For notation that is standard in a given subject area, judgement is needed to decide whether the definition should be given. Potential confusion can often be avoided by using redundant words. For example, if $\rho(A)$ is not obviously the spectral radius of the matrix A you can say "the spectral radius $\rho(A)$".

The second question is "where?" The practice of giving a long sequence of definitions at the start of a work is not recommended. Ideally, a definition should be given in the place where the term being defined is first used. If it is given much earlier, the reader will have to refer back, with a possible loss of concentration (or worse, interest). Try to minimize the distance between a definition and its place of first use.

It is not uncommon for an author to forget to define a new term on its first occurrence. For example, Steenrod uses the term "grasshopper reader" on page 6 of his essay on mathematical writing [256], but does not define it until it occurs again on the next page.

To reinforce notation that has not been used for a few pages you may be able to use redundancy. For example, "The optimal steplength α^* can be found as follows." This implicit redefinition either reminds readers what α^* is, or reassures them that they have remembered it correctly.

Finally, how should a term be defined? There may be a unique definition or there may be several possibilities (a good example is the term M-matrix, which can be defined in at least fifty different ways [23]). You should aim for a definition that is short, expressed in terms of a fundamental property or idea, and consistent with related definitions. As an example, the standard definition of a normal matrix is a matrix $A \in \mathbb{C}^{n\times n}$ for which $A^*A = AA^*$ (where $*$ denotes the conjugate transpose). There are at least 70 different ways of characterizing normality [119], but none has the simplicity and ease of use of the condition $A^*A = AA^*$.

By convention, *if* means *if and only if* in definitions, so do not write "The graph G is connected if and only if there is a path from every node in G to every other node in G." Write "The graph G is connected if there is a path from every node in G to every other node in G" (and note that this definition can be rewritten to omit the symbol G). It is common practice to italicize the word that is being defined: "A graph is *connected* if there is a path from every node to every other node." This has the advantage of making it perfectly clear that a definition is being given, and not a result. This emphasis can also be imparted by writing "A graph is defined to be connected if ...", or "A graph is said to be connected if ...."

If you have not done so before, it is instructive to study the definitions in a good dictionary. They display many of the attributes of a good mathematical definition: they are concise, precise, consistent with other definitions, and easy to understand.

Definitions of symbols are usually made with a simple equality, perhaps preceded by the word "let" if they are in-line, as in "let $q(x) = ax^2+bx+c$." Various other notations have been devised to give emphasis to a definition,

including

$$q(x) := ax^2 + bx + c,$$
$$ax^2 + bx + c =: q(x),$$
$$q(x) \stackrel{\text{def}}{=} ax^2 + bx + c,$$
$$q(x) \equiv ax^2 + bx + c,$$
$$q(x) \triangleq ax^2 + bx + c.$$

If you use one of these special notations you must use it consistently, otherwise the reader may not know whether a straightforward equality is meant to be a definition.

3.5. Notation

Consider the following extract.

> Let $\widehat{H_k} = Q_k^H \widetilde{H_k} Q_k$, partition $X = [X_1, X_2]$ and let $\mathcal{X} = \text{range}(X_1)$. Let U^* denote the nearest orthonormal matrix to X_1 in the 2-norm.

These two sentences are full of potentially confusing notation. The distinction between the hat and the tilde in $\widehat{H_k}$ and $\widetilde{H_k}$ is slight enough to make these symbols difficult to distinguish. The symbols $\mathcal{X}$ and X are also too similar for easy recognition. Given that $\mathcal{X}$ is used, it would be more consistent to give it a subscript 1. The name H_k is unfortunate, because H is being used to denote the conjugate transpose, and it might be necessary to refer to $\widetilde{H_k}^H$! Since A^* is a standard synonym for A^H, the use of a superscripted asterisk to denote optimality is confusing.

As this example shows, the choice of notation deserves careful thought. Good notation strikes a balance among the possibly conflicting aims of being readable, natural, conventional, concise, logical and aesthetically pleasing. As with definitions, the amount of notation should be minimized.

Although there are 26 letters in the alphabet and nearly as many again in the Greek alphabet, our choice diminishes rapidly when we consider existing connotations. Traditionally, ϵ and δ denote small quantities, i, j, k, m and n are integers (or i or j the imaginary unit), λ is an eigenvalue and π and e are fundamental constants; π is also used to denote a permutation. These conventions should be respected. But by modifying and combining eligible letters we widen our choice. Thus γ and A yield, for example, $\widehat{A}$, $\overline{A}$, $\widetilde{A}$, A', γ_A, A_γ, $\mathbf{A}$, $\mathcal{A}$, $\mathbb{A}$.

Particular areas of mathematics have their own notational conventions. For example, in numerical linear algebra lower case Greek letters represent

scalars, lower case roman letters represent column vectors, and upper case Greek or roman letters represent matrices. This convention was introduced by Householder [143].

In his book on the symmetric eigenvalue problem [217], Parlett uses the symmetric letters A, H, M, T, U, V, W, X, Y to denote symmetric matrices and the symmetric Greek letters Λ, Θ, Φ, Δ to denote diagonal matrices. Actually, the roman letters printed above are not symmetric because they are slanted, but Parlett's book uses a sans serif mathematics font that yields the desired symmetry. Parlett uses this elegant, but restrictive, convention to good effect.

We can sometimes simplify an expression by giving a meaning to extreme cases of notation. Consider the display

$$\beta_{ij} = \begin{cases} 0, & i > j, \\ \dfrac{1}{u_j}, & i = j, \\ \dfrac{1}{u_j} \displaystyle\prod_{r=i}^{j-1} \left(\frac{-c_r}{u_r} \right), & i < j. \end{cases}$$

There are really only two cases: $i > j$ and $i \leq j$. This structure is reflected and the display made more compact if we define the empty product to be 1, and write

$$\beta_{ij} = \begin{cases} 0, & \text{if } i > j, \\ \dfrac{1}{u_j} \displaystyle\prod_{r=i}^{j-1} \left(\frac{-c_r}{u_r} \right), & \text{if } i \leq j. \end{cases}$$

(Here, I have put "if" before each condition, which is optional in this type of display.) Incidentally, note that in a matrix product the order of evaluation needs to be specified: $\prod_{i=1}^{n} A_i$ could mean $A_1 A_2 \ldots A_n$ or $A_n A_{n-1} \ldots A_1$.

Notation also plays a higher level role in affecting the way a method or proof is presented. For example, the $n \times n$ matrix multiplication $C = AB$ can be expressed in terms of scalars,

$$c_{ij} = \sum_{k=1}^{n} a_{ik} b_{kj}, \qquad 1 \leq i, j \leq n,$$

or at the matrix-vector level,

$$C = [\, Ab_1, \quad Ab_2, \quad \ldots, \quad Ab_n \,],$$

where $B = [b_1, b_2, \ldots, b_n]$ is a partition into columns. One of these two viewpoints may be superior, depending on the circumstances. A deeper

example is provided by the fast Fourier transform (FFT). The discrete Fourier transform (DFT) is a product $y = F_n x$, where F_n is the unitary Vandermonde matrix with (r, s) element $\omega^{(r-1)(s-1)}$ $(1 \le r, s \le n)$, and $\omega = \exp(-2\pi i/n)$. The FFT is a way of forming this product in $O(n \log n)$ operations. It is traditionally expressed through equations such as the following (copied from a numerical methods textbook):

$$\sum_{j=0}^{n-1} e^{2\pi ijk/n} f_j = \sum_{j=0}^{n/2-1} e^{2\pi ikj/(n/2)} f_{2j} + \omega^k \sum_{j=0}^{n/2-1} e^{2\pi ikj/(n/2)} f_{2j+1}.$$

The language of matrix factorizations can be used to give a higher level description. If $n = 2m$, the matrix F_n can be factorized as

$$F_n \Pi_n = \begin{bmatrix} I_m & \Omega_m \\ I_m & -\Omega_m \end{bmatrix} \begin{bmatrix} F_m & 0 \\ 0 & F_m \end{bmatrix},$$

where Π_n is a permutation matrix and $\Omega_m = \operatorname{diag}(1, \omega, \ldots, \omega^{m-1})$. This factorization shows that an n-point DFT can be computed from two $n/2$-point transforms, and this reduction is the gist of the radix-2 FFT. The book *Computational Frameworks for the Fast Fourier Transform* by Van Loan [284], from which this factorization is taken, shows how, by using matrix notation, the many variants of the FFT can be unified and made easier to understand.

An extended example of how notation can be improved is given by Gillman in the appendix titled "The Use of Symbols: A Case Study" of *Writing Mathematics Well* [104]. Gillman takes the proof of a theorem by Sierpinski (1933) and shows how simplifying the notation leads to a better proof. Knuth set his students the task of simplifying Gillman's version even further, and four solutions are given in [164, §21].

Mathematicians are always searching for better notation. Knuth [163] describes two notations that he and his students have been using for many years and that he thinks deserve widespread adoption. One is notation for the Stirling numbers. The other is the notation $\mathcal{S}$, where $\mathcal{S}$ is any true-or-false statement. The definition is

$$[\mathcal{S}] = \begin{cases} 1, & \text{if } \mathcal{S} \text{ is true,} \\ 0, & \text{if } \mathcal{S} \text{ is false.} \end{cases}$$

For example, the Kronecker delta can be expressed as $\delta_{ij} = [i = j]$. The square bracket notation will seem natural to those who program; indeed, Knuth adapted it from a similar notation in the 1962 book by Iverson that led to the programming language APL [144, p. 11]. The square bracket notation is used in the textbook *Concrete Mathematics* [116]; that book

and Knuth's paper give a convincing demonstration of the usefulness of the notation.

Halmos has these words to say about two of his contributions to mathematical notation [127]:

> My most nearly immortal contributions to mathematics are an abbreviation and a typographical symbol. I invented "iff," for "if and only if"—but I could never believe that I was really its first inventor .... The symbol is definitely not my invention—it appeared in popular magazines (not mathematical ones) before I adopted it, but, once again, I seem to have introduced it into mathematics. It is the symbol that sometimes looks like □, and is used to indicate an end, usually the end of a proof. It is most frequently called the "tombstone," but at least one generous author referred to it as the "halmos".

Table 3.1 shows the date of first use in print of some standard symbols; some of them are not as old as you might expect. Not all these notations met with approval when they were introduced. In 1842 Augustus de Morgan complained (quoted by Cajori [49, p. 328 (Vol. II)]):

> Among the worst of barbarisms is that of introducing symbols which are quite new in mathematical, but perfectly understood in common, language. Writers have borrowed from the Germans the abbreviation $n!$ to signify $1.2.3.\ldots.(n-1)n$, which gives their pages the appearance of expressing surprise and admiration that 2, 3, 4, etc., should be found in mathematical results.

3.6. Words versus Symbols

Mathematicians are supposed to like numbers and symbols, but I think many of us prefer words. If we had to choose between reading a paper dominated by symbols and one dominated by words then, all other things being equal, most of us would choose the wordy paper, because we would expect it to be easier to understand. One of the decisions constantly facing the mathematical writer is how to express ideas: in symbols, in words, or both. I suggest some guidelines.

- Use symbols if the idea would be too cumbersome to express in words, or if it is important to make a precise mathematical statement.
- Use words as long as they do not take up much more space than the corresponding symbols.

Table 3.1. First use in print of some symbols. Sources: [49], [116], [163].

Symbol	Name	Year of publication
∞	infinity	1655 (Wallis)
π	pi (3.14159...)	1706 (Jones)
e	e (2.71828...)	1736 (Euler)
i	imaginary unit ($\sqrt{-1}$)	1794 (Euler)
$\equiv$	congruence	1801 (Gauss)
$n!$	factorial	1808 (Kramp)
$\sum$	summation	1820 (Fourier)
$\binom{n}{k}$	binomial coefficient	1826 (von Ettinghausen)
$\prod$	product	1829 (Jacobi)
∇	nabla	1853 (Hamilton)
δ_{ij}	Kronecker delta	1868 (Kronecker)
$\lvert z \rvert$	absolute value	1876 (Weierstrass)
$O(f(n))$	big oh	1894 (Bachmann)
$\lfloor x \rfloor$, $\lceil x \rceil$	floor, ceiling	1962 (Iverson)

- Explain in words what the symbols mean if you think the reader might have difficulty grasping the meaning or essential feature.

Here are some examples.

> (1) Define $C \in \mathbb{R}^{n \times n}$ by the property that $\mathrm{vec}(C)$ is the eigenvector corresponding to the smallest eigenvalue in magnitude of A, where the vec operator stacks the columns of a matrix into one long vector.

To make this definition using equations takes much more space, and is not worthwhile unless the notation that needs to be introduced (in this case, a name for $\min\{\, |\lambda| : \lambda \text{ is an eigenvalue of } A \,\}$) is used elsewhere. A possible objection to the above wordy definition of vec is that it does not specify in which order the columns are stacked, but that can be overcome by appending "taking the columns in order from first to last".

> (2) Since $|g'(0)| > 1$, zero is a repelling fixed point, so x_k does not tend to zero as $k \to \infty$.

An alternative is

> Since $|g'(0)| > 1$, 0 is a repelling fixed point, so $x_k \not\to 0$ as $k \to \infty$.

This sentence is only slightly shorter than the original and is harder to read—the symbols are beginning to intrude on the grammatical structure of the sentence.

> (3) If $B \in \mathbb{R}^{n\times n}$ has a unique eigenvalue λ of largest modulus then $B^k \approx \lambda^k x y^T$, where $Bx = \lambda x$ and $y^T B = \lambda y^T$ with $y^T x = 1$.

The alternative of "where x and y are a right and left eigenvector corresponding to λ, respectively, and $y^T x = 1$" is cumbersome.

> (4) Under these conditions the perturbed least squares solution $x + \Delta x$ can be shown to satisfy
>
> $$\frac{\|\Delta x\|_2}{\|x\|_2} \le \epsilon\kappa_2(A)\left(1 + \frac{\|b\|_2}{\|A\|_2\|x\|_2}\right) + \epsilon\,\kappa_2(A)^2 \frac{\|r\|_2}{\|A\|_2\|x\|_2} + O(\epsilon^2).$$
>
> Thus the sensitivity of x is measured by $\kappa_2(A)$ if the residual r is zero or small, and otherwise by $\kappa_2(A)^2$.

Here, we have a complicated bound that demands an explanation in words, lest the reader overlook the significant role played by the residual r.

> (5) If $y_1, y_2, \ldots, y_n$ are all $\neq 1$ then $g(y_1, y_2, \ldots, y_n) > 0$.

In the first sentence "all $\neq 1$" is a clumsy juxtaposition of word and equation and most writers would express the statement differently. Possibilities include

> If $y_i \neq 1$ for $i = 1, 2, \ldots, n$, then $g(y_1, y_2, \ldots, y_n) > 0$.
> If none of the y_i $(i = 1, 2, \ldots, n)$ equals 1, then $g(y_1, y_2, \ldots, y_n) > 0$.

If the condition were "$\neq 0$" instead of "$\neq 1$", then it could simply be replaced by the word "nonzero". In cases such as this, the choice between words and symbols in the text (as opposed to in displayed equations) is a matter of taste; good taste is acquired by reading a lot of well-written mathematics.

The symbols $\forall$ and $\exists$ are widely used in handwritten notes and are an intrinsic part of the language in logic. But generally, in equations that are in-line, they are better replaced by the equivalent words "for all" and

"there exists". In displayed equations either the symbol or the phrase is acceptable, though I usually prefer the phrase. Compare

$$\sigma\big(G(t)\big) = \exp\big(t\sigma(A)\big) \quad \text{for all } t \geq 0$$

with

$$\sigma\big(G(t)\big) = \exp\big(t\sigma(A)\big) \quad \forall t \geq 0.$$

Similar comments apply to the symbols $\Rightarrow$ (implies) and $\Longleftrightarrow$ (if and only if), though these symbols are more common in displayed formulas.

Of course, for some standard phrases that appear in displayed formulas, there is no equivalent symbol:

$$\text{minimize} \quad c^T x - \mu \sum_{i=1}^{n} \ln x_i \quad \text{subject to } Ax = b,$$

$$\underline{\lim}\, \tfrac{1}{n} \log D_j^n < 0 \quad \text{almost surely,}$$

$$z^T y = \|z\|_D \|y\| = 1, \quad \text{where} \quad \|z\|_D = \max_{v \neq 0} \frac{|z^T v|}{\|v\|}.$$

3.7. Displaying Equations

An equation is displayed when it needs to be numbered, when it would be hard to read if placed in-line, or when it merits special attention, perhaps because it contains the first occurrence of an important variable. The following extract gives an illustration of what and what not to display.

> Because $\delta(\bar{x}, \mu)$ is the smallest value of $\|\overline{X}z/\mu - e\|$ for all vectors y and z satisfying $A^T y + z = c$, we have
>
> $$\delta(\bar{x}, \mu) \leq \|\tfrac{1}{\mu}\overline{X}z - e\|.$$
>
> Using the relations $z = \mu X^{-1} s$ and $\bar{x}_i = 2x_i - x_i s_i$ gives
>
> $$\frac{1}{\mu}\overline{X}z = \overline{X}X^{-1}s = (2X - XS)X^{-1}s = 2s - S^2 e.$$
>
> Therefore, $\delta(\bar{x}, \mu) \leq \|2s - S^2 e - e\|$, which means that

$$\delta(\bar{x}, \mu)^2 \leq \sum_{i=1}^{n} (2s_i - s_i^2 - 1)^2 = \sum_{i=1}^{n} (s_i - 1)^4 \leq \left(\sum_{i=1}^{n} (s_i - 1)^2 \right)^2 = \delta(x, \mu)^4.$$

> The condition $\delta(x, \mu) < 1$ thus ensures that the Newton iterates $\overline{x}$ converge quadratically.

The second and third displayed equations are too complicated to put in-line. The first $\delta(\overline{x}, \mu)$ inequality is displayed because it is used in conjunction with the second display and it is helpful to the reader to display both these steps of the argument. The consequent inequality $\delta(\overline{x}, \mu) \leq \|2s - S^2 e - e\|$ fits nicely in-line, and since it is used immediately it is not necessary to display it.

When a displayed formula is too long to fit on one line it should be broken before a binary operation. Example:

$$|e_{m+1}| \leq |G^{m+1} e_0| + c_n u(1 + \theta_x)\{c(A)|(I - G)^D M^{-1}| + (m+1)|(I - E)M^{-1}|\}(|M| + |N|)|x|.$$

The indentation on the second line should take the continuation expression past the beginning of the left operand of the binary operation at which the break occurred, though, as this example illustrates, this is not always possible for long expressions. A formula in the text should be broken after a relation symbol or binary operation symbol, not before.

3.8. Parallelism

Parallelism should be used, where appropriate, to aid readability and understanding. Consider this extract:

> The Cayley transform is defined by $C = (A - \theta_1 I)^{-1}(A - \theta_2 I)$. If λ is an eigenvalue of A then
>
> $$(\lambda - \theta_2)(\lambda - \theta_1)^{-1}$$
>
> is an eigenvalue of C.

The factors in the eigenvalue expression are presented in the reverse order to the factors in the expression for C. This may confuse the reader, who might, at first, think there is an error. The two expressions should be ordered in the same way.

Parallelism works at many levels, from equations and sentences to theorem statements and section headings. It should be borne in mind throughout the writing process. If one theorem is very similar to another, the statements should reflect that—the wording should not be changed just for the sake of variety (see *elegant variation*, §4.15). However, it is perfectly acceptable to economize on words by saying, in Theorem 2 (say) "Under the conditions of Theorem 1".

For a more subtle example, consider the sentence

It is easy to see that $f(x, y) > 0$ for $x > y$.

In words, this sentence is read as "It is easy to see that $f(x, y)$ is greater than zero for x greater than y." The first $>$ translates to "is greater than" and the second to "greater than", so there is a lack of parallelism, which the reader may find disturbing. A simple cure is to rewrite the sentence:

It is easy to see that $f(x, y) > 0$ when $x > y$.
It is easy to see that if $x > y$ then $f(x, y) > 0$.

3.9. Dos and Don'ts of Mathematical Writing

Punctuating Expressions

Mathematical expressions are part of the sentence and so should be punctuated. In the following display, all the punctuation marks are necessary. (The second displayed equation might be better moved in-line.)

The three most commonly used matrix norms in numerical analysis are particular cases of the Hölder p-norm

$$\|A\|_p = \max_{x \neq 0} \frac{\|Ax\|_p}{\|x\|_p}, \qquad A \in \mathbb{R}^{m \times n},$$

where $p \geq 1$ and

$$\|x\|_p = \Big(\sum_{i=1}^{n} |x_i|^p \Big)^{1/p}.$$

Otiose Symbols

Do not use mathematical symbols unless they serve a purpose. In the sentence "A symmetric positive definite matrix A has real eigenvalues" there is no need to name the matrix unless the name is used in a following sentence. Similarly, in the sentence "This algorithm has $t = \log_2 n$ stages", the "$t =$ " can be omitted unless t is defined in this sentence and used immediately. Watch out for unnecessary parentheses, as in the phrase "the matrix $(A - \lambda I)$ is singular."

Placement of Symbols

Avoid starting a sentence with a mathematical expression, particularly if a previous sentence ended with one, otherwise the reader may have difficulty parsing the sentence. For example, "A is an ill-conditioned matrix"

(possible confusion with the word "A") can be changed to "The matrix A is ill-conditioned."

Separate mathematical symbols by punctuation marks or words, if possible, for the same reason.

Bad: If $x > 1$ $f(x) < 0$.
Fair: If $x > 1$, $f(x) < 0$.
Good: If $x > 1$ then $f(x) < 0$.

Bad: Since $p^{-1} + q^{-1} = 1$, $\|\cdot\|_p$ and $\|\cdot\|_q$ are dual norms.
Good: Since $p^{-1} + q^{-1} = 1$, the norms $\|\cdot\|_p$ and $\|\cdot\|_q$ are dual.

Bad: It suffices to show that $\|H\|_p = n^{1/p}$, $1 \le p \le 2$.
Good: It suffices to show that $\|H\|_p = n^{1/p}$ for $1 \le p \le 2$.
Good: It suffices to show that $\|H\|_p = n^{1/p}$ $(1 \le p \le 2)$.

Bad: For $n = r$ (2.2) holds with $\delta_r = 0$.
Good: For $n = r$, (2.2) holds with $\delta_r = 0$.
Good: For $n = r$, inequality (2.2) holds with $\delta_r = 0$.

"The" or "A"

In mathematical writing the use of the article "the" can be inappropriate when the object to which it refers is (potentially) not unique or does not exist. Rewording, or changing the article to "a", usually solves the problem.

Bad: Let the Schur decomposition of A be QTQ^*.
Good: Let a Schur decomposition of A be QTQ^*.
Good: Let A have the Schur decomposition QTQ^*.

Bad: Under what conditions does the iteration converge to the solution of $f(x) = 0$?
Good: Under what conditions does the iteration converge to a solution of $f(x) = 0$?

Notational Synonyms

Sometimes you have a choice of notational synonyms, one of which is preferable. In the following examples, the first of each pair is, to me, the more aesthetically pleasing or easier to read (a capital letter denotes a matrix).

$$\Big(\sum_{i,j}(a_{ij}-b_{ij})^2\Big)^{1/2}, \qquad \sqrt{\sum_{i,j}(a_{i,j}-b_{i,j})^2},$$

$$\exp\big(2\pi i(x^2+y^2)^{-1/2}\big), \qquad e^{\frac{2\pi i}{\sqrt{x^2+y^2}}},$$

$$(1-n\epsilon)^{-1}|L||U|, \qquad \frac{|L||U|}{1-n\epsilon},$$

$$X_{k+1} = \tfrac{1}{2}X_k(3I-X_k^2), \qquad X_{k+1} = \frac{X_k}{2}[3I-X_k^2],$$

$$\min\{\,\epsilon : |b-Ay| \le \epsilon|A||y|\,\}, \qquad \min\{\,\epsilon \mid |b-Ay| \le \epsilon|A||y|\,\},$$

$$\min\{\,\|A-UBP\| : U^TU = I,\ P \text{ a permutation}\,\},$$

$$\min_{\substack{U^TU=I \\ P \text{ a permutation}}} \|A-UBP\|.$$

In the next two examples, the first form is preferable because it saves space without a loss of readability.

$$x = [\,x_1,\ \ x_2,\ \ \ldots,\ \ x_n\,]^T, \qquad x = \begin{bmatrix} x_1 \\ x_2 \\ \vdots \\ x_n \end{bmatrix},$$

$$\Lambda = \operatorname{diag}(\lambda_i), \qquad \Lambda = \begin{bmatrix} \lambda_1 & & & \\ & \lambda_2 & & \\ & & \ddots & \\ & & & \lambda_n \end{bmatrix}.$$

Of course, the diag(·) notation should be defined if it is not regarded as standard.

Referencing Equations

When you reference an earlier equation it helps the reader if you add a word or phrase describing the nature of that equation. The aim is to save the reader the trouble of turning back to look at the earlier equation. For example, "From the definition (6.2) of dual norm" is more helpful than "From (6.2)"; and "Combining the recurrence (3.14) with inequality (2.9)" is more helpful than "Combining (3.14) and (2.9)". Mermin [200] calls this advice the "Good Samaritan Rule". As in these examples, the word added should be something more informative than just "equation" (or the ugly abbreviation "Eq."), and inequalities, implications and lone expressions should not be referred to as equations.

Miscellaneous

When working with complex numbers it is best not to use "i" as a counting index, to avoid confusion with the imaginary unit. More generally, do not use a letter as a dummy variable if it is already being used for another purpose.

Note the difference between the Greek letter epsilon, ϵ, and the "belongs to" symbol $\in$, as in $\|x\| \leq \epsilon$ and $x \in \mathbb{R}^n$. Another version of the Greek epsilon is ε. Note the distinction between the Greek letter π and the product symbol $\prod$.

By convention, standard mathematical functions such as sin, cos, arctan, max, gcd, trace and lim are set in roman type, as are multiple-letter variable names. It is a common mistake to set these in italic type, which is ambiguous. For example, is $tanx$ the product of four scalars or the tangent of x?

In bracketing multilayered expressions you have a choice of brackets for the layers and a choice of sizes, for example $\{\,[\,(\{[($, this ordering being the one recommended by *The Chicago Manual of Style* [58]. Most authors try to avoid mixing different brackets in the same expression, as it leads to a rather muddled appearance.

Write "the kth term", not "the k^{th} term", "the k'th term" or "the k-th term." (It is interesting to note that nth is a genuine word that can be found in most dictionaries.)

A slashed exponent, as in $y^{1/2}$, is generally preferable to a stacked one, as in $y^{\frac{1}{2}}$.

The standard way to express that i is to take the values 1 to n in steps of 1 is to write

$$i = 1, \ldots, n \quad \text{or} \quad i = 1, 2, \ldots, n,$$

where all the commas are required. An alternative notation originating in programming languages such as Fortran 90 and MATLAB is $i = 1{:}n$. For counting down we can write $i = n, n-1, \ldots, 1$ or $i = n{:}-1{:}1$, where the middle integer denotes the increment. This notation is particularly convenient when extended to describe submatrices: $A(i{:}j, p{:}q)$ denotes the submatrix formed from the intersection of rows i to j and columns p to q of the matrix A.

Avoid (or rewrite) tall in-line expressions, such as $\begin{bmatrix} g_1 \\ g_2 \end{bmatrix}$, which can disrupt the line spacing.

There are two different kinds of ellipsis: vertically centred ($\cdots$) and "ground level" or "baseline" ($\ldots$). Generally, the former is used between operators such as $+$, $=$, and $\leq$, and the latter is used between a list of

Glossary for Mathematical Writing

1. Without loss of generality = I have done an easy special case.
2. By a straightforward computation = I lost my notes.
3. The details are left to the reader = I can't do it.
4. The following alternative proof of X's result may be of interest = I cannot understand X.
5. It will be observed that = I hope you hadn't noticed that.
6. Correct to within an order of magnitude = wrong.

Adapted from [222].

symbols or to indicate a product. Examples:

$$x_1 + x_2 + \cdots + x_n, \quad \sigma_1 \geq \sigma_2 \geq \cdots \geq \sigma_n, \quad \lambda_1, \lambda_2, \ldots, \lambda_n, \quad A_1 A_2 \ldots A_n.$$

An operator or comma should be symmetrically placed around the ellipsis; thus $x_1 + x_2 + \cdots x_n$ and $\lambda_1, \lambda_2, \ldots \lambda_n$ are incorrect.

When an ellipsis falls at the end of a sentence there is the question of how the full stop (or period) is treated. Recommendations vary. *The Chicago Manual of Style* suggests typing the full stop before the three ellipsis points (so that there is no space between the first of the four dots and the preceding character). When the ellipsis is part of a mathematical formula it seems natural to put it before the full stop, but the two possibilities may be visually indistinguishable, as in the sentence

> The Mandelbrot set is defined in terms of the iteration $z_{k+1} = z_k^2 + c$, $k = 0, 1, 2, \ldots$.

A vertically centred dot is useful for denoting multiplication in expressions where terms need to be separated for clarity:

$$16046641 = 13 \cdot 37 \cdot 73 \cdot 457,$$

$$\operatorname{cond}(A, x) = \frac{\| \, |I - A^+ A| \cdot |A^T| \cdot |A^{+^T} x| \, \|}{\|x\|}.$$

Care is needed to avoid ambiguity in slashed fractions. For example, the expression $-(b-a)^3/12 f''(\eta)$ is better written as $-\big((b-a)^3/12\big) f''(\eta)$ or $-f''(\eta)(b-a)^3/12$.