How to write mathematics

Effective mathematical communication is an important skill, yet is rarely taught. Effective communication is important for three reasons:

1. It helps others (e.g. your peers, your teachers, your markers) to understand and appreciate your ideas.
2. It enables *you* to understand your ideas, six months later, when you can't remember what you were thinking when you wrote them.
3. It enhances your ability to communicate in nonmathematical spheres:
   - To clearly express scientific ideas.
   - To write good computer code and clear documentation or specifications for such code.
   - To formulate clear and compelling arguments in any subject.

Effective mathematical communication is based on several principles.

Words vs. symbols

Mathematical prose relies heavily on symbols, but it is important not to use symbols to excess. Effective communication requires *complete, grammatically correct sentences*. Symbolism should never replace sentences.

Mathematical symbols are used to represent complex formulae or longwinded phrases which appear repeatedly in the text. By introducing a new symbol, the author avoids constantly rewriting these formulae (and the reader avoids constantly rereading them). When used judiciously in this fashion, symbolism can enhance clarity. Mathematical symbols should *only* be used in this way.

Introducing new symbols can simplify the exposition, but at a cost: it increases the burden on the reader's memory. Only introduce new symbols if the benefit outweighs the cost.

Any sequence of symbols should read like a fragment of an English sentence (eg. '$a \in \mathbf{B} \subset \mathbf{C}$' means '$a$ is an element of $\mathbf{B}$, which is a subset of $\mathbf{C}$'). Thus, mathematical symbols should be surrounded by enough English words to form a complete sentence. If a mathematical assertion is nothing but symbols, it is entirely unreadable.

Most important are the stock phrases which indicate the *logical role* which an assertion plays in your argument. For example:

Hypotheticals: 'Assume that...', 'Suppose that...', 'Let...', etc.

Consequences: 'It follows that...', 'Thus...', 'Therefore....', 'But then....', '...so...', etc.

Logical Connectives: '...and...', '...but...', 'either....or...', 'if....then....otherwise....', 'Since.....it follows....', etc.

Contradistinctions: 'On the other hand...', 'However...', etc.

Reminders: 'Recall...', 'Observe that...', 'Note...', 'Now...', 'Remark', etc.

Desiderata: 'We want to show that...', 'I claim that...', etc.

Conclusions: 'We conclude that...',

Appropriate use of these words can greatly clarify an argument. Omitting them can utterly obscure it. A common mistake of novices is to write a 'proof' consisting of a sequence of naked mathematical assertions, with not a single word explaining their logical role. This is impossible to read: it is impossible to know which assertions are *hypotheses*, which are *consequences* of these hypotheses, and which are *conclusions*.

Equalities

Many novices have a bad habit of 'proving' an equality by writing the equality *first*, and then 'working backwards'. While this is sometimes an effective problem-solving strategy, it is *not a correct proof*. It gives the reader the impression that you are assuming what you are trying to prove. For example, suppose we want to prove the trigonometric identity $\sec^2(\theta) \;=\; 1+\tan^2(\theta)$....

Bad:

$$\begin{aligned} \sec^2(\theta) &= 1+\tan^2(\theta) \\ \frac{1}{\cos^2(\theta)} &= 1+\frac{\sin^2(\theta)}{\cos^2(\theta)} \\ \frac{1}{\cos^2(\theta)} &= \frac{\cos^2(\theta)}{\cos^2(\theta)}+\frac{\sin^2(\theta)}{\cos^2(\theta)} \\ \frac{1}{\cos^2(\theta)} &= \frac{\cos^2(\theta)+\sin^2(\theta)}{\cos^2(\theta)} \\ 1 &= \cos^2(\theta)+\sin^2(\theta) \\ 1 &= 1 \end{aligned}$$

Good:

$$\begin{aligned} \sec^2(\theta) &= \frac{1}{\cos^2(\theta)} \\ &= \frac{\cos^2(\theta)+\sin^2(\theta)}{\cos^2(\theta)} \\ &= \frac{\cos^2(\theta)}{\cos^2(\theta)}+\frac{\sin^2(\theta)}{\cos^2(\theta)} \\ &= 1+\frac{\sin^2(\theta)}{\cos^2(\theta)} \\ &= 1+\tan^2(\theta) \end{aligned}$$

Modules

A mathematical proof should be broken down into distinct *modules*, each of which solves a particular problem or accomplishes a particular goal. These modules are comparable to *subroutines* within a computer program.

Each module should *begin* by clearly stating its goal. Avoid the 'abracadabra' writing style, where you first present a confusing mass of technicalities, and then at the *end*, explain what it was all about. For example:

Bad:
bleah bleah... Thus we see that all frobnitzes have the type II Siegel property.

Good:
I claim that that all frobnitzes have the type II Siegel property. To see this, observe that bleah.

A module should have a logical development like a proper English essay: distinct paragraphs, each developing a distinct idea, and each logically flowing into the next.

Often, a proof has a *hierarchical* structure, with submodules nested within modules. This hierarchical structure should be *explicitly visible* in the page layout. For example:

Bad:

I claim that that all frobnitzes have the type II Siegel property. To see this, first we must show that frobnitzes are spurling. Bleah bleah bleah bleah bleah bleah bleah bleah bleah bleah bleah bleah bleah bleah bleah bleah bleah bleah bleah bleah. Next I claim that spurling implies the bi-infinite receding foobaz condition. bleah. Finally, note that the foobaz condition implies the Siegel property: bleah.

Good:

Claim 1: *All frobnitzes have the type II Siegel property.*

Proof:

Claim 1.1: *Frobnitzes are spurling.*

Proof: Bleah bleah. ..... □ [Claim 1.1]

Claim 1.2: *Spurling implies the bi-infinite receding foobaz condition.*

Proof: bleah bleah bleah bleah bleah bleah bleah bleah bleah bleah bleahbleah bleah bleah bleah bleah bleah bleah bleah bleah bleah bleah bleah bleah bleah. ..... □ [Claim 1.2]

Claim 1.3: *The Foobaz condition implies the Siegel property*

Proof: bleah. ..... □ [Claim 1.3]

From Claims 1.1, 1.2, and 1.3, we conclude that frobnitzes have the Siegel property □ [Claim 1]

INTUITION PUMPS

At the beginning, sketch out the strategy of your proof in broad, intuitive terms. This sketch need not be rigorous, precise, or even mathematically correct (though you must clearly acknowledge where you bend the truth). The sketch should involve as little notation as possible. It should provide the reader with a rough 'mental framework', upon which to 'attach' the technicalities of of the proof. Generally, there are only two reasons why you would be unwilling to provide such a sketch:

1. You don't want to reveal your 'secret insight', which enabled you to solve the problem.
2. You don't really understand what you've done... you just cobbled together a bunch of machinery, and it all works, somehow.

Neither of these will endear you to the reader.

NOTATION Good notation is crucial for effective communication, and depends upon three principles: *Simplicity*, *Mimesis*, and *Consistency*.

I. Notational Simplicity: *Notation should be as simple as possible, while conveying all essential information.* Avoid attaching extra subscripts, superscripts, tildes, hats, primes, arguments, or other dongles, unless they are *strictly necessary* to avoid ambiguity.

Sometimes, you must choose between:

(A) Notation which is *technically correct*, but horribly complex and confusing.

(B) Notation which is *technically incorrect*, but much simpler to read, and whose meaning is obvious to anyone with half a brain.

In such a situation, you should always choose **(B)**, but you must *explicitly point out* that you are using a 'technically incorrect' notation. This is sometimes called 'abusing notation'. Some common examples:

- If $\mathbf{X}$ and $\mathbf{Y}$ are two sets (groups, topological spaces, etc.) and $f : \mathbf{X} \longrightarrow \mathbf{Y}$ is an injection that embeds $\mathbf{X}$ as a subset (subgroup, subspace, etc.) of $\mathbf{Y}$, then we sometimes identify $\mathbf{X}$ with its image $f(\mathbf{X})$, and identify each point $x \in \mathbf{X}$ with its image $f(x) \in f(\mathbf{X})$.

- In analysis, one often sees arguments involving sequences, subsequences, subsubsequences, etc. The standard 'reindexing' trick allows one to avoid introducing subsubscripts, subsubsubscripts, etc. For example:

Bad:
Let $\{x_n\}_{n=1}^{\infty}$ be a sequence in $\mathbf{X}$. Since $\mathbf{X}$ has the type II Siegel property, we can find a subsequence $\{x_{n_i}\}_{i=1}^{\infty}$ satisfying the bi-infinite receding foobaz condition. Next, by Gromwald's Other Theorem, we can find a subsubsequence $\{x_{n_{i_j}}\}_{j=1}^{\infty}$ which is strongly frobnitz. But then Lemma 18(c) yields a subsubsubsequence $\{x_{n_{i_{j_k}}}\}_{k=1}^{\infty}$ converging to a fixed point.

Good:
Let $\{x_n\}_{n=1}^{\infty}$ be a sequence in $\mathbf{X}$. Now, $\mathbf{X}$ has the type II Siegel property; so by dropping to a subsequence and reindexing, we can assume that $\{x_n\}_{n=1}^{\infty}$ satisfies the bi-infinite receding foobaz condition. Next, by Gromwald's Other Theorem, we can drop to a subsequence which is strongly frobnitz. By Lemma 18(c), we can drop to a further subsequence and reindex, and assume that $\{x_n\}_{n=1}^{\infty}$ converges to a fixed point.

II. Notational Mimesis: *Notational relationships should reflect mathematical relationships.* This mnemonic device helps the reader remember the meanings of myriad symbols. In particular:

1. Use a specific *font* to denote objects of a particular *type*. For example, in linear algebra, you might denote *vectors* by bold-faced lower-case letters ($\mathbf{u}, \mathbf{v}, \mathbf{w}, \ldots$), *matrices* by bold-faced *upper*-case letters ($\mathbf{A}, \mathbf{B}, \mathbf{C}, \ldots$), *scalars* by lower-case roman letters ($r, s, t, \ldots$), and vector (sub)spaces by upper-case 'blackboard' font ($\mathbb{U}, \mathbb{V}, \mathbb{W}, \ldots$).

2. Use letters which stand for descriptive words. Hence, 'f' stands for 'function'; 'n' means 'number'; 'p' means 'prime'; '$\mathbf{S}$' stands for 'set'; '$\mathbf{G}$' stands for 'group', etc.

3. Use *alphabetically consecutive* letters to denote similar objects. For example, if one function is called 'f', then the next two functions could be 'g' and 'h'.

   Problem: One alphabetical sequence may run into another. For example, if functions are $f, g, h, \ldots$, and indexes are $i, j, k, \ldots$, then one can't represent more than three different functions.

4. Use letters from the same *lexicographical family* to denote objects which 'belong' together. For example:

   (a) If $\mathbf{S}$ and $\mathbf{T}$ are sets, then elements of $\mathbf{S}$ should be $s_1, s_2, s_3, \ldots$, while elements of $\mathbf{T}$ are $t_1, t_2, t_3, \ldots$.

(b) Use indexing variables which agree with their bounds. Reserve upper-case letters (eg. $J, K, L, M, N, \ldots$) for the bounds of intervals or indexing sets, and then use the corresponding lower-case letters (eg. $j, k, l, m, n, \ldots$) as indexes. For example:

Bad:
For all $r \in [1..n]$ and $q \in [1..m]$, let
$$A(r,q) \;=\; \sum_{i=1}^{k} \sum_{j=1}^{\ell} a_{ij}(r,q).$$

Good:
For all $n \in [1..N]$ and $m \in [1..M]$, let
$$A(n,m) \;=\; \sum_{j=1}^{J} \sum_{k=1}^{K} a_{jk}(n,m).$$

(c) If $\mathbf{v}$ is a vector, then its entries should be called $v_1, \ldots, v_N$. If $\mathbf{A}$ is a matrix, then its entries should be $a_{11}, \ldots, a_{NM}$. If $\mathbb{V}$ is a vector space, then elements of $\mathbb{V}$ should be $\mathbf{v}_1, \mathbf{v}_2, \mathbf{v}_3$.

Problem: The entries of vector $\mathbf{v}_1$ would be denoted with cumbersome double-subscripts $v_{11}, v_{12}, \ldots, v_{1N}$. There are three solutions:

i. Distinguish elements of $\mathbb{V}$ with 'decorations', eg. $\mathbf{v}, \widetilde{\mathbf{v}}, \widehat{\mathbf{v}}, \mathbf{v}', \overline{\mathbf{v}}, \mathbf{v}^*, \mathbf{v}^\dagger$, etc. Thus, entries of $\widetilde{\mathbf{v}}$ would be denoted $\widetilde{v}_1, \widetilde{v}_2, \ldots, \widetilde{v}_N$, etc.

Problem: This is impractical when there are more than six or seven distinct elements. Also, some decorations have special meanings in some contexts. For example, $\mathbf{v}'$ might mean 'derivative', $\overline{\mathbf{v}}$ might be 'complex-conjugate'; $\mathbf{v}^*$ might be 'adjoint'; $\widehat{\mathbf{v}}$ might be 'Fourier transform'; etc.

ii. Distinguish elements of $\mathbb{V}$ with *superscripts*, eg. $\mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \ldots$. Thus, entries of $\mathbf{v}^{(1)}$ would be $v_1^{(1)}, v_2^{(1)}, \ldots, v_N^{(1)}$. This allows you to represent any number of elements (including infinite sequences),

Problem: The 'double indexing' can be very confusing. Notation with multiple indices is hard to read, and should be avoided when possible.

iii. Use *alphabetically consecutive* letters, eg. $\mathbf{u}, \mathbf{v}, \mathbf{w}, \ldots$. Thus, entries of $\mathbf{u}$ would be $u_1, \ldots, u_N$. We've already mentioned the shortcomings of this method.

Letters from the same lexicographical family should not be used to excess. For example, a proof involving vectors $\mathbf{v}, \mathbf{v}^{(1)}, \mathbf{v}^{(2)}, \mathbf{v}^{(3)}, \widetilde{\mathbf{v}}, \widetilde{\mathbf{v}}^{(1)}, \widetilde{\mathbf{v}}^{(2)}, \mathbf{v}', \widetilde{\mathbf{v}}', \widehat{\mathbf{v}}$, and $\overline{\mathbf{v}}$ can become very confusing; it would be better to use distinct letters $\mathbf{u}, \mathbf{v}, \mathbf{w}, \ldots$.

III. Notational Consistency: *The same notational conventions should be used throughout the text.* Decide at the outset what conventions you will use, and stick to them. For example, if, on page 1, you use 'f', 'g', and 'h' for functions, and '$\mathbf{U}$', '$\mathbf{V}$', and '$\mathbf{W}$' for open sets, then on page 10, do not suddenly switch to 'ϕ', 'ψ', and 'ξ' for your functions and '$\mathcal{X}$', '$\mathcal{Y}$', and '$\mathcal{Z}$' for sets.

Whenever possible, conform to the notational conventions established in prior literature. If everyone else uses 'ϕ' for to mean a morphism and '$\mathbf{K}$' to mean its kernel, do not insist on using 'f' and '$\mathbf{X}$' instead, unless you have a very good reason. On the other hand, do not slavishly adhere to stupid conventions that could clearly be improved.

Pictures

A common misconception is that pictures are 'unmathematical' or 'nonrigorous', and should be replaced by symbolism whenever possible. In fact, the opposite is true. Most mathematics is motivated by visual intuitions, and most mathematicians think visually. A page of symbolism is an extremely inadequate substitute for a few good pictures. Math books

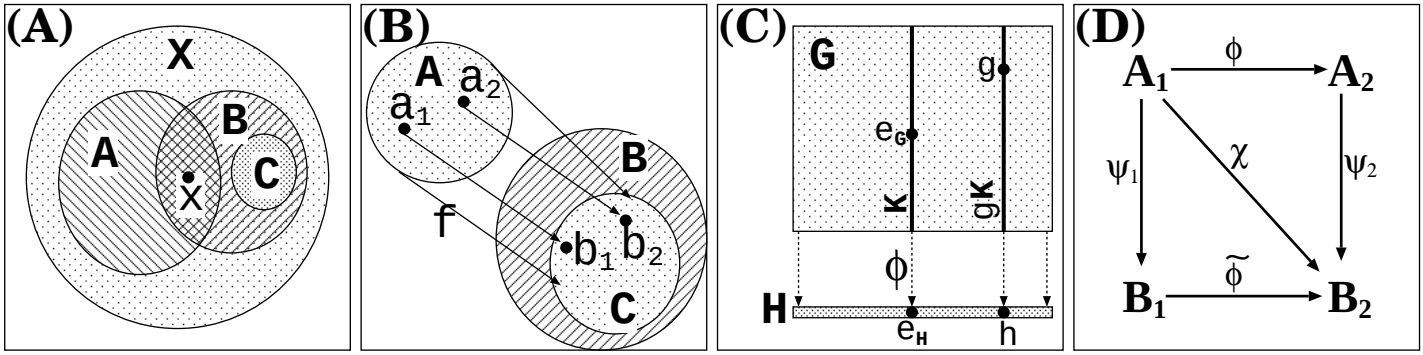

Figure 1:

should be filled with illustrations —at least one per page. They aren't because publishers are cheap.

Pictures can never replace a clear written exposition; a proof still needs words. But pictures can *always* enhance the clarity of the exposition. Some areas of mathematics (eg. calculus, linear algebra, geometry, topology) are explicitly geometric in nature, and the value of pictures is obvious. But pictures are also useful in 'abstract' mathematics...

- ...to depict relationships between sets. For example, Figure 1A shows how $x \in \mathbf{A} \cap \mathbf{B}$, $\mathbf{C} \subset \mathbf{B} \setminus \mathbf{A}$, and $\mathbf{A} \cup \mathbf{B} \subset \mathbf{X}$. This is called a *Venn diagram*.

- ...to depict functions. For example, Figure 1B shows that f is a function from $\mathbf{A}$ into $\mathbf{B}$, with image $\mathbf{C}$. Also, $f(a_1) = b_1$, and $f(a_2) = b_2$.

- ...to depict algebraic structure, by treating groups, rings, etc. as metaphorical 'vector spaces'. For example, Figure 1C shows that $\phi : \mathbf{G} \longrightarrow \mathbf{H}$ is a homomorphism with kernel $\mathbf{K}$. Here, $\phi(g) = h$, and the preimage of h is the coset $g\mathbf{K}$.

- ...to depict networks of functions between spaces. For example, Figure 1D depicts functions $\phi : \mathbf{A}_1 \longrightarrow \mathbf{A}_2$, $\psi_1 : \mathbf{A}_1 \longrightarrow \mathbf{B}_1$, $\widetilde{\phi} : \mathbf{B}_1 \longrightarrow \mathbf{B}_2$, $\psi_2 : \mathbf{A}_2 \longrightarrow \mathbf{B}_2$, and $\chi : \mathbf{A}_1 \longrightarrow \mathbf{B}_2$, such that $\psi_2 \circ \phi = \chi = \widetilde{\phi} \circ \psi_1$. This is called a *commuting diagram*.