\diamond \diamond \diamond

Natural T_EX Notation in Mathematics

MICHAL MARVAN*

ABSTRACT. In this paper we introduce Nath, a $IAT_EX 2.09/2_{\varepsilon}$ style implementing a natural TEX notation for mathematics.

KEYWORDS: Natural mathematical notation

The paradigm

The original idea behind T_EX was to overcome the principal contradiction of scientific typography, namely that typographers shape publications without understanding their content. But now that the field has been conquered and T_EX has become a standard in scientific communication we face an unwanted effect: a decline in typographic quality.

The LATEX scheme of separating the presentation and content (in the sty and doc files, respectively) already enabled a basic division of responsibility between scientists and typographers, with a positive impact on the quality of both professional and lay publishing. In contemporary LATEX we have a rather firmly established content markup for the main parts of a text such as sections, lists, theorems; thanks to styles from the American Mathematical Society we also have a wide variety of mathematical symbols and environments. But the notation for mathematical expressions still encodes presentation.

The basic mathematical constructs of plain T_EX refer to presentation by the obvious requirement of universality, and the same holds true for their later reencodings. For instance, \frac , which differs from plain T_EX 's \over only by syntax, is a presentation markup to the effect that two elements are positioned one above the other, centered, and separated by a horizontal line. Even though T_EX has no technical difficulty typesetting built-up fractions (such as $\frac{A}{B}$) in text style, publishers that still adhere to fine typography may prefer converting them to the slash form A/B. The conversion, during which the mathematical content must not be changed, cannot be reliably done by nonexperts. However, the chance that software can perform the conversion does not turn out to be completely unrealistic, as we shall see below.

Namely, herewith we would like to introduce a *natural mathematical notation* in T_EX . By such we mean the coarsest notation for which there exist algorithms that find a typographically sound and mathematically correct context-dependent presentation,

*Mathematical Institute, Silesian University in Opava, Bezručovo nám. 13, 746 01 Opava, Czech Republic. *E-mail:* Michal.Marvan@math.slu.cz

whereas the notation itself is essentially independent of the context. It should be stressed that it is not the goal of natural notation to encode the mathematical content – in contrast to the notation used, e.g., in computer algebra or programming languages.

An accompanying $I_{TEX} 2_{\epsilon}$ style, Nath, is offered to the public for scrutiny of its strengths and weaknesses, and for practical use. Its mission is to produce traditional mathematical typography with minimal requirements for presentation markup. The price is that T_{EX} spends more time on formatting the mathematical material. However, savings in human work appear to be substantial, the benefits being even more obvious in the context of lay publishing, when expert guidance is an exception.

Preliminaries

Nowadays we recognize two major styles to typeset mathematical material, which we shall call display and in-line. They possibly occur in three sizes: text, script and second level script. The *display* style is characterized by the employment of vertical constructions and arbitrarily sizeable brackets, with emphasis on legibility. Over centuries the display style was commonplace even in in-line formulas, forcing typographers

to spread lines as with $\lim_{n\to\infty} \left(1+\frac{1}{n}\right)^n$. Irregular line spacing and extra costs due to

unused white space were among the arguments pushed forward against this practice.

Gradually the *in-line* style evolved, essentially within the Royal Society of London [3, p. 275]. Based on suggestions by the famous logician Augustus de Morgan, the style was introduced by G.G. Stokes [18] and gained strong support from J.W. Rayleigh [15] (all three gentlemen were presidents of the Society). Designed for use under strict limits on the vertical size, the in-line style replaces stacking with horizontal linking, as in $\lim_{n\to\infty} (1+1/n)^n$. Typical is the use of the solidus "/" as a replacement for the horizontal bar of a fraction. The in-line style adds some more ambiguity to that already present in the display style.

Accordingly, Nath uses two distinct math modes, display and in-line, which are fundamentally distinct in TEXnical aspects and go far beyond plain TEX's four math styles (\displaystyle, \textstyle, \scriptstyle, and \scriptscriptstyle). The single dollar sign \$ starts the in-line mode; otherwise said, Nath's in-line formulas use in-line mode, and so do sub- and superscripts even within displayed formulas. The double dollar sign \$\$ as well as various display math environments start display mode. In contrast to TEX's defaults but in good agreement with old traditions, Nath's default mode for numerators and denominators of displayed fractions is display.

Either mathematical mode can be forced on virtually any subexpression by making it an argument of one of the newly introduced \displayed and \inline commands. Preserved for backward compatibility, plain T_EX's \displaystyle and \textstyle only affect the size of type, like \scriptstyle and \scriptstyle . Actually, no such simple switch can alter the fairly sophisticated Nath mode.

OPERATORS

We start with a solution to a subtle problem that occurs in both the display and inline styles, namely, uneven spacing around symbols of binary operations following an operator, as in λ id -g. Recall that T_FX's capability of producing fine mathematical typography depends on the assignment of one of the eight types (Ord, Op, Bin, Rel, Open, Close, Punct, Inner) to every math atom (see [7, pp. 158 and 170]). Oddly enough, [7, Appendix G, rule 5] says that a Bin atom (a binary operation) preceded by an Op (an operator) becomes an Ord (an ordinary atom like any variable). However, the existence of expressions like $\lambda \operatorname{id} - g$ suggests that operators followed by binary operations make perfect sense. Therefore, we propose that the spacing between a Bin atom preceded by an Op be a medium space, i.e., the value in the 2nd row and 3rd column of the table on p. 170 of the TFXbook [7] be '(2)' instead of '*'. Since TFX provides us with no means to change the table, we had to redefine \mathop to a "mixed-type creator," namely \mathop from the left and \mathord from the right, augmented with appropriate handling of exceptions when Op's behaviour differs from that of Ord. Fortunately, the exceptions occur only when the following atom is Open (an opening delimiter) or Punct (a punctuation), which can be easily recognized by comparison to a fairly short lists of existing left delimiters and punctuation marks. One only must successfully pass over sub- and superscripts as well as over the \limits and \nolimits modifiers that may follow the Op atom, which on the positive side gives us an opportunity to enable \setminus in Op's subscripts, so that

\$\$

\sum_{i,j \in K \\ i \ne j} a_{ij} \$\$

prints as

$$\sum_{\substack{i,j\in K\\i\neq j}} a_{ij}.$$

Another new mixed-type object is !, which produces suitable spacing around factorials: (m!n!) typesets as (m!n!). It is simply the exclamation mark (which itself is of type Close) with $mathopen{}\mathinner{}\$ appended from the right.

Nath also supports a handy notation for abbreviations in a mathematical formula, such as $e^{2\pi i} = -1$, $ad_x y$, $span\{u, v\}$, $H' = H'_{symm} + H'_{antisymm}$, $f|_{int U}$. They are created as letter strings starting from a back quote, e.g., $f(e^{2\pi i})$, $f(ad_x y)$, etc.

Fractions

There are three basic types of fractions in modern scientific typography:

1) built-up:
$$\frac{A}{B}$$
, 2) piece: $\frac{1}{2}$, 3) solidus: A/B

Mostly they indicate division in a very broad sense; often but not always they can be recast in an alternative form, such as AB^{-1} or A : B (e.g., $\partial f / \partial x$ cannot). Type 1

fractions are now restricted exclusively to display style. The solidus form is mandatory for non-numeric fractions in in-line style; it is also spontaneously preferred in specific situations such as quotient algebraic structures (e.g., $\mathbf{Z}/2\mathbf{Z}$). Type 2 is strictly confined to numeric fractions (when the numerator and denominator are explicit decimal numbers), e.g., $\frac{5}{7}$, $\frac{1}{10000}$, $\frac{0.15}{1.22}$. Numeric fractions should be of type 2 or 3 in in-line style. In display style they may occur as both types 1 and 2, depending on the vertical size of the adjacent material. When one changes from display to in-line style, built-up fractions are generally substituted with solidus fractions, and parentheses may have to be added to preserve the mathematical meaning.

Nath supports two commands to typeset fractions: slash / and \frac. With slash one always typesets a type 3 fraction. With \frac one creates a fraction whose type is determined by the following rules: In display style, non-numeric fractions come out as type 1. The type of numeric fractions is determined by the *principle of smallest fences*: A numeric fraction is typeset as a built-up fraction in display style if and only if this will not extend any paired delimiter present in the expression. We explicitly discontinue the tradition according to which numeric fractions adjust their size to the next symbol. For example, Nath typesets

```
$$ (\int rac 12 + x)^2 - (\int rac 12 + \int rac 1x)^2
$$ as
```

 $\left(\frac{1}{2}+x\right)^2 - \left(\frac{1}{2}+\frac{1}{x}\right)^2.$

In the sequel, we shall need some definitions. A symbol is said to be *exposed* if it is neither enclosed in paired delimiters nor contained in a sub- or superscript nor used in a construction with an extended line (such as \sqrt, \overline, or a wide accent). Next, by an element of type Bin* we shall mean an element that either is of type Bin or is of type Ord, starts an expression, and originates from an element of type Bin by [7, Appendix G, rule 5].

In in-line style, the rules that govern typesetting of $\exists AB$ are as follows. If A, B are numeric (i.e., strings composed of decimal numbers, spaces and decimal points), then the resulting fraction is of type 2. Otherwise the fraction is of type 3 and bracketing subroutines are invoked. Parentheses are put around the numerator A if A contains an exposed element of type Bin, Rel, Op; or an exposed delimiter that is not a paired delimiter (e.g., $/, \setminus \text{ or } |$). Likewise, parentheses are put around the denominator B if B contains an exposed element of type Bin^{*}, Rel; or an exposed delimiter that is not a paired delimiter. Finally, parentheses are put around the whole fraction if at least one of the columns of Table 1 contains 'Yes' in the corresponding row. For example,

$$\frac{a+\frac{b}{b+c}}{1-c} \longrightarrow (a+b/(b+c))/(1-c).$$

Type	Left neighbour	Example	Right neighbour	Example
Ord	Yes^1	x(a/b)	Yes	(a/b)x
Op	Yes	$\sin(a/b)$	Yes	$(a/b)\sin x$
Bin^*	No^2	1 + a/b	No	a/b + 1
Rel	No	=a/b	No	a/b =
Open	No	[a/b]	Yes	(a/b)[
Close	Yes](a/b)	No	a/b]
Punct	No	,a/b	No	a/b,
Inner	Yes^1	$\frac{1}{2}a/b$	Yes	$(a/b)^{\frac{1}{2}}$

¹ No, if the left neighbour is a digit or a piece fraction (hence Inner) and at the same time A starts with neither Bin^{*} nor digit nor decimal point. E.g., $\frac{1}{2}a/b$, but $\frac{1}{2}(-2a/b)$, $\frac{1}{2}(25a/b)$, $\frac{1}{2}(.5a/b)$. ² Yes, if A starts with Bin^{*}, e.g., 1 + (-a/b).

TABLE 1: BRACKETING RULES FOR FRACTIONS

Nath's approach to binary operations is mathematically correct under the following assumption: Every binary operator * that occurs in the numerator, denominator, or the immediate vicinity of \frac is, similarly to addition, of lower precedence than /. An obvious exception is the multiplication " \cdot ", which is, however, left associative with respect to division and hence $A \cdot B/C = A \cdot (B/C) = (A \cdot B)/C$ anyway. (We also assume that numerators and denominators do not contain exposed punctuation, except for the decimal point.) In particular, Nath converts

$$\frac{A}{B} \otimes \frac{C}{D} \qquad \rightarrow \qquad A/B \otimes C/D,$$

and

$$\frac{A \otimes B}{C \otimes D} \longrightarrow (A \otimes B)/(C \otimes D).$$

Literature contains examples of different bracketing, $(A/B) \otimes (B/C)$ and $A \otimes B/C \otimes D$, namely $H_n((K/C) \otimes L)$ in [11, Ch. V, diag. 10.6] and Ker $\partial_n/C_n \otimes A$ in [11, Ch. V, proof of Th. 11.1]. Anyway, we feel that giving more binding power to ' \otimes ' than to '/' is unfounded.

Now we come to a more delicate question, which reflects a difference between human readability and machine readability. In favour of the former it is often desirable to suppress unneeded parentheses; compare $\exp(x/2\pi)$ and $\exp(x/(2\pi))$. This is one of the reasons why Nath converts

$$\frac{a}{bc} \longrightarrow a/bc$$

and not a/(bc). Here we follow the living tradition according to which a/bc means 'a divided by bc.' Numerous examples of use in professional publications can be easily documented, e.g., [2, p. 9, 34, 52, 89, 115], and the convention is by no means outdated, see, e.g., [14]. It is supported by major manuals of style, albeit by means of examples,

such as $|\langle X_1, X_2 \rangle| / ||X_1|| ||X_2||$ in [4]. As much as one chapter in Wick's handbook [20] is devoted to solidus fractions and examples of use; all of them use the same rule as above.

Unfortunately, the convention is not completely devoid of controversy. Some opponents argue that if bc means multiplication, then a/bc = (a/b)c by the current standard rules of precedence, and therefore one should write a/(bc) to have both b and c in the denominator. But, examples like x/12, $\partial f/\partial x \partial y$, 1/f(x), $1/\sin x$, $\mathbb{Z}/2\mathbb{Z}$ show that not every juxtaposition denotes multiplication, while in all of these cases an added pair of parentheses would be certainly superfluous. Understanding juxtaposition requires understanding mathematics, for which reason it is certainly preferable that typography treats all juxtapositions on an essentially equal footing (an exception being subtle rules for close and loose juxtapositions, see the expression $\sin xy \cos xy$ in [1]).

The core of the problem resides in the possible ambiguity of the juxtaposition (see Fateman and Caspi [8], who bring lots of examples of ambiguous notation in the context of machine recognition of TEX-encoded mathematics). However, we feel that by all reasonable criteria, the ambiguity should be kept limited within the denominator, instead of letting it propagate beyond the fraction, which is exactly what would happen if we adapted the competitive rule a/bc = (a/b)c. Indeed, the mathematical interpretation of juxtaposition is context dependent, a good case in point being the classic a(x+y). Its meaning depends on whether a is a function that may have x + y as its argument, or not. Under Nath's rules 1/a(x + y) is invariably equal to 1/(a(x + y)), while under the competitive rule the meaning of 1/a(x + y) would be (1/a)(x + y) in case of a = const! But then we conclude that the traditional rule a/bc = a/(bc) remains the only reasonable alternative for an unthinking machine.

Anyway, we must admit that there is currently no general consent on this point. The AIP style manual [1] says: "do not write 1/3x unless you mean 1/(3x)," while the Royal Statistical Society [16] considers the notation a/bc "ambiguous if used without a special convention." The Annals of Mathematical Statistics even changed its rules from $1/2\pi$ to $1/(2\pi)$ between 1970 and 1971. Use of programming languages and symbolic algebra systems with different syntactic rules also has a confusing effect.

It is certainly true that ambiguity of notation makes reading of mathematical publications more difficult than absolutely necessary. A good solution, which is heartily recommended, amounts to typesetting all difficult fractions in display, or disambiguating them through explicit use of parentheses or otherwise.

Nath's solution is, we believe, the best possible from those available, given the fact that T_EX does not provide tools for recognizing close (spaceless) juxtaposition. Nath essentially treats juxtaposition as a binary operation of higher precedence than solidus (even the loose juxtaposition expressed via a small amount of white space, such as \timmuskip around Op's):

$$\frac{1}{\cos x} \longrightarrow 1/\cos x.$$

The only exception is that the right binding power of loose juxtaposition is considered uncomparable to the left binding power of the solidus, so that, e.g., $\sin x/y$ comes out

Left delimiters		Right delimiters	
(())
[,\lbrack	[],\rbrack]
$\{, \lbrace$	{	$\}, $ rbrace	}
<, \langle	<	>, \rangle	\rangle
\lfloor	L	\rfloor	
\lceil	Γ	\rceil]
lvert, left		\rvert, \right	
\lBrack, \double[[\rBrack, \double]]]
\lAngle, \double<	$\langle\!\langle$	\rAngle, \double>	$\rangle\rangle$
\lFloor	L	\rFloor	
\lCeil	Π	\rCeil	Π
\lVert, \ldouble		\rvert, \rdouble	
\triple[\triple]]]]
\triple<	///	\triple>	$\rangle\rangle\rangle$
\ltriple		\rtriple	

TABLE 2: PAIRED DELIMITERS

as truly ambiguous (following [1]); hence Nath converts

 $\sin \frac{x}{y} \longrightarrow \sin(x/y)$

and

$$\frac{\sin x}{y} \longrightarrow (\sin x)/y$$

– even though Wick interprets $\sin x/y$ as $(\sin x)/y$.

Delimiters

Plain T_EX introduces various delimiter modifiers such as \left and \right. If used continually without actual need, as is often done, they produce unsatisfactory results; such continual use is as undesirable as is the failure to use them when they are actually needed. Under natural notation every left parenthesis is a left delimiter by default, and Nath does its best to ensure proper matching to whatever is enclosed.

Table 2 lists paired delimiters. Their presentation depends on the current mode. In display mode delimiters automatically adjust their size and positioning to match the material enclosed (thus rendering \left and \right nearly obsolete), and do so across line breaks (which themselves are indicated by mere \\ whenever allowed by the context). Around asymmetric formulas the delimiters may be positioned asymmetrically.

A particularly nice example, taken from [9, p. 4], is

$$\frac{M}{\left(1 - \frac{x_1 + \dots + x_n + pZ}{r}\right) \left(1 - p\frac{\frac{\partial Z}{\partial x_2} + \dots + \frac{\partial Z}{\partial x_n}}{\rho}\right)}$$

(no modifiers in front of the parentheses).

The modifiers **\double** and **\triple** create double and triple counterparts of delimiters, such as

$$\left[\left[\frac{x}{y} \right] \right]$$

We also introduce *middle delimiters*: $\mid and \middle | produce | , \Mid and \double | produce || , and \triple | produces |||. They have exactly the size of the nearest outer pair of delimiters. For example:$

$$\Big\{(x_i) \in R^\infty \,\Big| \, \sum_{i=1}^\infty x_i^2 = 1 \Big\}.$$

Observe that matching is done in a subtle way, disregarding sub- and superscripts, accents, and other negligible parts. (Let us also note that in order to implement the above-mentioned principle of smallest fences in display style, Nath represents numeric fractions as middle delimiters.)

With nested delimiters it is frequently desirable that the outer delimiters are bigger than the inner ones. In displayed formulas this is controlled by a count \delimgrowth that when set to n makes every nth delimiter bigger. One should set $\delimgrowth=1$ when a display contains many vertical bars:

$$C_6 \left| \left| f \int_{\Omega} \left| \tilde{S}_{a,-}^{-1,0} W_2(\Omega,\Gamma_1) \right| \right| \left| \left| u \right| \to W_2^{\tilde{A}}(\Omega;\Gamma_r,T) \right| \right|.$$

(cf. [17]).

In in-line mode a completely different mechanism is needed, which would be applicable to implicit delimiters introduced by frac. We introduce a *command* big having the effect that the next entered level of delimiters will be set in big size (in plain T_EX 's sense). For instance, $\beta \ge frac 1{f(x)}$ produces $\Delta(1/f(x))$. It is an error to place a big within delimiters that are not big. Observe that Nath's big need not immediately precede a delimiter; this gave us an opportunity to introduce bigg as an abbreviation for bigbig.

Unbalanced delimiters may be present in an in-line formula (as is often the case in indices in differential geometry), but then cannot be resized.

DISPLAYED FORMULAS

Displayed formulas have never been a serious problem. Yet there is room for innovation and simplification of the current presentation markup. Downes presented a style [6], which breaks multiline displayed equations automatically. With Nath, every end of line must be marked by an explicit $\$, but this $\$ can be used at almost any place where it makes sense. In particular, \$ $\dots = \dots \$ is a valid syntax. The result is a multiline formula without special alignment:

(by default, Nath indents all displayed equations by \mathindent=4pc). Within the equation environment, the formula obtains a single centered number.

A kind of alignment can be obtained with the wall-return construction. The syntax is \wall ... \\ ... \\ ... \return, and can be nested. Here is an example \$\$

The meaning is that the typeset material must not cross the wall.

Display delimiters cannot be combined with alignments unless every cell has balanced delimiters, which is certainly the case with matrices, but not necessarily with other alignment environments, such as eqnarray. The purpose of these environments is, essentially, to align binary relation symbols in one of the two typographically relevant situations:

(1) an *n*-line sequence of equalities and relations;

(2) a pile of n distinct formulas.

In case (1), walls alone represent a simple alternative that spares the 2n alignment symbols & required by equarray. It is also possible to put a wall-return block into one cell of an alignment.

Acknowledgements

I would like to thank many people, explicitly Eylon Caspi, Michael Guravage, Wietze van der Laan, Petr Sojka, Simon Pepping, Alex Verbovetsky, and staff in the National Library of Scotland in Edinburgh.

REFERENCES

- [1] AIP Style Manual, 4th edition (Amer. Inst. Physics, New York, 1990).
- [2] S.H. Benton, The Hamilton-Jacobi Equation. A Global Approach (Academic Press, New York, 1977).
- [3] F. Cajori, A History of Mathematical Notations I, II (Open Court, Chicago, 1928, 1929).
- [4] The Chicago Manual of Style (The Univ. Chicago Press, Chicago and London, 1969).
- [5] A. De Morgan, The calculus of functions, in: Encyclopaedia Metropolitana (1845).
- [6] M. Downes, Breaking equations, TUGboat 18 (1997) 182–194.
- [7] D.E. Knuth, The T_EXbook (Addison Wesley, Reading, 1984)
- [8] R.J. Fateman and E. Caspi, Parsing T_EX into mathematics, poster presented at International Symposium on Symbolic and Algebraic Computation (ISSAC '99), Vancouver BC Canada, July 28–31, 1999.
- [9] E. Goursat, Leşons sur l'Intégration des Équations aux Dérivées Partielles du Premier Ordre, 2nd ed. (Hermann, Paris, 1921)
- [10] J. Larmor, Practical suggestions on mathematical notation and printing, an appendix to [15].
- [11] S. Mac Lane, *Homology* (Springer, Berlin, 1967).
- [12] A. Macfarlane, Sir George Gabriel Stokes, in: Lectures on Ten British Physicists of the Nineteenth Century (New York, 1919).
- [13] M. Marvan, Přirozená matematická notace v T_EXu, in: J. Kasprzak and P. Sojka, eds., *SLT 2001* (Konvoj, Brno, 2001) 53–59.
- [14] Mathematical Composition Setters, Sample style setting, online www.mcs-ltd.co.uk, accessed Jan. 8, 2001.
- [15] J.W. Rayleigh, Address of the president, Lord Rayleigh, O.M., D.C.L., at the anniversary meeting on November 30, 1918, Proc. Roy. Soc. London, Sect. A 82 (1909) 1–17.
- [16] Royal Statistical Society, Notes for the preparation of mathematical papers, J. Roy. Stat. Soc., Ser. A 136 (1973) 293–294.
- [17] A. Samarin, F. Mittelbach and M. Goossens, The LATEX Companion (Addison-Wesley, Reading, 1994).
- [18] G.G. Stokes, Mathematical and Physical Papers, Vol. I (Cambridge Univ. Press, Cambridge, 1880).
- [19] E. Swanson, *Mathematics into Type* (Amer. Math. Soc., Providence, 1987).
- [20] K. Wick, *Rules for Typesetting Mathematics* (Mouton, The Hague, 1965); translated from the Czech original *Pravidla Matematické Sazby* (Academia, Praha, 1964).