Single- and multi-letter identifiers in Unicode mathematics

Will Robertson

Abstract

This paper argues that separate spacing and therefore separate font invocations is needed for single- and multi-letter identifiers in mathematical contexts. This is not explicitly provided for by Unicode mathematics nor by the OpenType mathematics fonts that currently exist. The unicode-math package has been extended to accommodate these needs, which provides better compatibility with legacy LaTeX and amsmath documents.

1 Introduction

Unicode mathematics has been developing over the last fifteen years, spearheaded by Barbara Beeton [1, 2, 3], and is now approaching mainstream use. Ulrik Vieth has described OpenType mathematics and its relationship to TeX [6, 7]. Although work began on Unicode mathematics in XeTeX and LuaTeX several years prior [4], the unicode-math package was first formally released in 2010 [5].

Although much care was taken to incorporate as many mathematical alphabetic symbols into Unicode as possible, within the limitations imposed by the Unicode Technical Committee at the time, there is not a direct correspondence between the de facto mathematics alphabets defined in LaTeX and those used in Unicode. The original version of the unicode-math package did not take these differences into account, which has led to certain problems as Unicode-aware TeX engines have become more popular. This article will discuss the limitations of the original unicode-math package and the new interfaces set up to allow a smoother transition from legacy LaTeX documents to Unicode.

2 Brief overview

The unicode-math package builds on LaTeX's math font selection system and implements the entire symbol repertoire of Unicode mathematics (many thousands of glyphs) for direct use within LaTeX. As unicode-math requires the XeTeX or LuaTeX engine, glyphs can be inserted as Literal Unicode characters if desired. Control sequences are also defined in order to be able to access each symbol by name, based on the work of Barbara Beeton. Mathematical alphabet commands are also defined to emulate LaTeX's and amsmath's traditional `\mathbf`, `\mathfrak`, etc., commands. The difference with such alphabets in Unicode mathematics is that all glyphs come from within a single font, whereas in 8-bit TeX systems accessing these glyphs involved *switching* fonts of different styles.

$$\mathbf{G} + \mathbf{\Gamma} = G + \Gamma$$

```
\setmathfont{xits-math.otf}%
  [math-style=TeX]
```

```
$\symbf{G}+\symbf{\Gamma} = G+\Gamma$
```

Example 1: An example of 'TeX'-style mathematics.

$$\boldsymbol{G} + \boldsymbol{\Gamma} = G + \mathit{\Gamma}$$

```
\setmathfont{xits-math.otf}%
  [math-style=ISO]
```

```
$\symbf{G}+\symbf{\Gamma} = G+\Gamma$
```

Example 2: An example of 'ISO'-style mathematics.

A Unicode mathematics font is loaded with the command such as:

```
\setmathfont{xits-math.otf}
```

Optional arguments to the package or to the font loading command can be used to change the style of the mathematics; for example, publishers differ on whether Greek letters should be upright or italic by default, or whether bold Latin letters should be upright or italic. These cases are compared in Examples 1 and 2. These examples also show the use of the `\symbf` command to turn a single-letter alphabetic symbol 'bold' according to the prevailing style. This is a new command that replaces the use of `\mathbf` in previous versions of the package, for reasons that we will now discuss.

3 Single-letter vs. multi-letter symbols

Unicode mathematics defines a large number of mathematical alphabet symbol ranges, including alphabets of upright, italic, sans serif, fraktur, and script shapes in regular and bold weights, among others. Many of these support both Latin and Greek letters, plus a few assorted 'Greek-like' symbols such as `\nabla` (∇) and `\partial` (∂).

In legacy LaTeX and amsmath-based mathematics, the number of alphabets is more restricted, and in many cases commands such as `\mathbf` are used for both single-letter symbols and multi-letter identifies. In contrast, the design of Unicode mathematics fonts to date has including kerning around the alphabetic symbols suitable for single letter symbols only, and these symbols are often not appropriate to use in multi-letter contexts.

Examples are given following of mathematics in which single-letter and multi-letter identifiers are used in the different mathematical contexts for a range of alphabet styles.

3.1 Upright regular text

Upright regular-weight roman text is occasionally used in applied mathematics to represent constants such as e, the exponential:

$$\mathrm{e}^{-st} = \mathrm{e}^{-i\omega t} .$$

In some cases upright roman is used for the complex number ($\mathrm{i} = \sqrt{-1}$). Since these symbols may be used in contexts in which they appear adjacent to a separate symbol they should have additional spacing to separate them, such as in the (trivial) case of '$\mathrm{e}\,\mathrm{e} = \mathrm{e}^2$'. As another common example, the derivative operator $\frac{d}{dt}$ is often written $\frac{\mathrm{d}}{\mathrm{d}t}$.

In fluid dynamics in particular, dimensionless values such as the Reynolds number Re are often typeset using upright symbols:

$$\mathrm{Re} = \rho \mathbf{v} l / \mu .$$

Other dimensionless values include, for example, the Froude number (Fr) and the Strouhal number (St). Since these identifiers are essentially abbreviations of words and names, they should be kerned as text to bind the letters together visually.

3.2 Italic text

The aforementioned dimensionless numbers are also frequently typeset as italic text, instead of upright (roman) text:

$$\mathit{Re} = \rho \mathbf{v} l / \mu .$$

In such cases it is plain that there needs to be a clear distinction between adjacent symbols and the dimensionless number. For 'R' and 'e' there is little difference, but take the Froude number:

$$\mathit{Fr} \neq Fr$$

This is typeset with '`\mathit{Fr} \neq Fr`'. The extra kerning on the right hand side distinguishes the symbols as being separate.

Other examples from the literature include italic being used for operator-like notation, such as in these arbitrary cases taken from the literature:

$$\overline{\mathit{Sep}}_{\mathrm{I}}(\Sigma_1^{[C]}) , \qquad \frac{\partial |\mathit{det}\, g|}{\partial g_{\mu\nu}} .$$

3.3 Bold upright text

Bold upright alphabetic symbols are commonly used in physics and engineering to denote vectors and matrices:

$$\mathbf{x} = [x,\, \dot{x},\, \theta,\, \dot{\theta}]^{\mathrm{T}} .$$

In mathematics, it is also common for bold upright roman to be used for both single *and* multi-letter abbreviations such as

$$DG = \hom(G, \mathbf{R}/\mathbf{Z}) \qquad \mathbf{Grp} \to \mathbf{Ab} \to \mathbf{Grp}$$

In the first case, the symbols should be kerned as separate glyphs, whereas in the second the letters should be spaced as in text.

Further examples could be given for sans serif regular weight and bold, for example. In contrast, alphabet ranges such as fraktur ($\mathfrak{ABC}$) and script ($\mathcal{ABC}$) are generally used for denoting single symbols only — although it is possible to find examples of notation such as

$$\mathcal{H}\!om(\mathscr{H}, \mathscr{E}) .$$

The term '$\mathcal{H}\!om$' is in a calligraphic style, whereas its arguments are shown in a 'curly' script style; Unicode mathematics, for now, only supports the latter, but unicode-math can be configured to support both if the font (such as STIX) contains glyphs for both.

The various examples listed above using single- and multi-letter identifiers cannot be represented using a single Unicode mathematics font, since each alphabetic range has a fixed spacing; in general, to date all Unicode mathematics fonts space the alphabetic glyphs as individual letters. Furthermore, when used in 'word-like' contexts, it is important to recognise that multilingual varieties of strings such as 'sin', 'cos', and 'tan' are possible, and that the fonts used to typeset such identifiers should support the entirety of Unicode, not just the restricted set of alphabetic symbols defined as 'mathematical'.

4 Additions to unicode-math

These factors lead to a definite tension between the commands defined originally in unicode-math, which mapped directly to Unicode mathematical alphabetic symbols only, and the commands that users expected from legacy LaTeX documents.

Due to the clear requirements of supporting single- and multi-letter identifiers, the approach now taken in unicode-math is to define a new set of commands, `\symbf`, `\symup`, `\symsf`, and so on, which switch to the Unicode mathematics *symbols* that can be used for single-letter identifiers. For multi-letter operators, LaTeX's traditional `\mathit`, `\mathbf`, etc., commands are retained as *font switches*, and do not perform any 'remapping' on their inputs. Fraktur and calligraphic alphabets, for example, are defined using `\symfrak` and `\symcal`, but familiar LaTeX commands `\mathfrak` and `\mathcal` are provided as synonyms for backwards compatibility.

Unless explicit fonts for `\mathbf` and so on are selected (see next), they are automatically selected

$$\mathrm{ημ}(\omega) = y/\rho$$

```
\setmainfont{Iwona-Regular.otf}
\setmathfont{texgyrepagella-math.otf}%
  [Scale=0.85]
```

```
$ \mathup{ημ}(ω) = y/ρ $
```

Example 3: Greek text and symbols in mathematics.

$$\rho_p = \text{Ш}_p * \rho$$

```
\setmathfont{texgyrepagella-math.otf}%
  [Scale=0.85]
\setmathfontface\mathcyr{Charter Roman}
```

```
$ ρ_p = \mathcyr{Ш}_p \ast ρ $
```

Example 4: Cyrillic Sha being used in mathematics.

from the default text fonts. This allows, for example, the case in Greek schoolbooks in which the sine rule might be written as shown in Example 3. By the way, `\mathup` is defined in unicode-math as a synonym for `\mathrm` for exactly such situations; we're not always typesetting 'RoMan' text any more.

Restoring the idea of the '`\mathXYZ`' commands being 'text fonts in mathematics mode' also provided the opportunity to restore the functionality of LaTeX's `\DeclareMathAlphabet`, and to define a new unicode-math interface to it. It is now possible to write, say,

```
\setmathfontface\mathittt
  {texgyrecursor-italic.otf}
```

to define the math font command `\mathittt{...}`, which selects (in this case) an italic typewriter font. This command can also be used to select fonts for the built-in commands `\mathbf`, `\mathsf`, etc.

The `\setmathfontface` command also provides the possibility of selecting particular fonts for typesetting characters in other scripts. For example, the Cyrillic glyph 'Sha' is often used to typeset the Dirac Comb Function, and this can now be supported easily as shown in Example 4.

5 Compatibility

After upgrading to version 0.8, many users will wish to adapt their old documents written in unicode-math to the new syntax from `\mathbf` to `\symbf` and similar. Package options `mathbf=sym` will rename `\mathbf` to have behaviour as in previous version, with similar options for `mathsf` and so on. The legacy commands for switching to a text font in math mode are renamed to `\mathtextbf`, `\mathtextsf`, etc.

A common suggestion is to implement a 'smart' version of `\mathbf` that analyses the number of characters in its arguments. Writing `\mathbf{abc}` could switch to `\mathtextbf`, and `\mathbf{A}` to `\symbf`. This interface would be convenient for Latin-based examples, but further consideration is required before deciding it's actually a good idea.

In such a 'smart' version of the command, a case such as `\mathbf{\alpha\alpha}` would result in mathematics symbols inside a command that is intended for typesetting text, which would not produce expected output. (Exactly as in the case now in regular LaTeX when the unsuspecting user writes `\mathbf{\alpha}`.) In the 'smart' approach, having `\mathbf{\alpha}` work 'as expected' but then having `\mathbf{\alpha\alpha}` 'fail' does not seem like a sensible approach.

6 Conclusion

The unicode-math package now contains a full suite of commands to select alphabet styles suitable for both single- and multi-letter identifiers. For ideal typesetting purposes, developers of OpenType mathematics fonts should provide font features that allow selection of kerning suitable for both mathematics (single-letter) and 'text' (multi-letter) purposes.

References

[1] Barbara Beeton. Unicode and math, a combination whose time has come — Finally! *TUGboat*, 21(3):176–185, September 2000. http://tug.org/TUGboat/tb21-3/tb68beet.pdf.

[2] Barbara Beeton. The STIX project — from Unicode to fonts. *TUGboat*, 28(3):299–304, 2007. http://tug.org/TUGboat/tb28-3/tb90beet.pdf.

[3] Barbara Beeton, Asmus Freytag, and Murray Sargent III. Unicode support for mathematics. Technical report, Unicode, Inc., 2015. http://www.unicode.org/reports/tr25.

[4] Jonathan Kew. XeTeX live. *TUGboat*, 29(1):146–150, 2007. http://tug.org/TUGboat/tb29-1/tb91kew.pdf.

[5] Will Robertson. Unicode mathematics in LaTeX: Advantages and challenges. *TUGboat*, 31(2):211–220, 2010. http://tug.org/TUGboat/tb31-2/tb98robertson.pdf.

[6] Ulrik Vieth. Do we need a 'Cork' math font encoding? *TUGboat*, 29(3):426–434, 2008. http://tug.org/TUGboat/tb29-3/tb93vieth.pdf.

[7] Ulrik Vieth. OpenType math illuminated. *TUGboat*, 30(1):22–31, 2009. http://tug.org/TUGboat/tb30-1/tb94vieth.pdf.

◇ Will Robertson
The University of Adelaide, SA, Australia
will dot robertson (at) latex-project dot org