MicroFloatingPoints.jl: providing very small IEEE 754-compliant floating-point types

Frédéric Goualard

The Journal of Open Source Software

MicroFloatingPoints.jl: providing very small IEEE 754-compliant floating-point types

Frédéric Goualard [1]

1 Nantes Université, École Centrale Nantes, CNRS, LS2N, UMR 6004, Nantes, France

DOI: 10.21105/joss.07050

Software

- Review
- Repository
- Archive

Editor: Daniel S. Katz

Reviewers:

- @matbesancon
- @dannys4
- @mkitti

Submitted: 23 July 2024
Published: 18 September 2024

Summary

The IEEE 754 standard defines the representation and the properties of the floating-point numbers used as surrogates for reals in computer programs. Most programming languages only support the 32-bit (`Float32`) and 64-bit (`Float64`) formats implemented in hardware. Machine learning, computer graphics, and numerical algorithms analysis all have a need for smaller formats, which are often neither supported in hardware, nor are they available as established types in programming languages. The `MicroFloatingPoints.jl` Julia library offers a parametric type that can be instantiated to compute with IEEE 754-compliant floating-point numbers with varying ranges and precisions (up to and including `Float32`). It also provides the programmer with various means to visualize what is computed.

Statement of need

Proving the properties of numerical algorithms involving floating-point numbers can be a very challenging task. Insight can often be gained by systematically executing the algorithm under study for all possible inputs. There are, however, too many values to consider with the classically available types `Float32` and `Float64`. Hence there is a need for libraries that offer many smaller IEEE 754-compliant types to play with. SIPE (Lefèvre, 2013), FloatX (Flegar et al., 2019), and CPFloat (Fasi & Mikaitis, 2023), to name a few, are such libraries. However, being written in languages such as C or C++, they lack the interactivity and tight integration with graphical facilities that can be obtained from using script languages such as Julia. `MicroFloatingPoints.jl` is a Julia library that fills this need by offering a parametric type, `Floatmu`, that can be instantiated to simulate in software small floating-point types: `Floatmu{8,23}` is a type using 8 bits to represent the exponent and 23 bits for the fractional part, which is equivalent to `Float32`; `Floatmu{8,7}` is equivalent to the Google Brain `bfloat16` format, ...

A quick tour

To obtain a (pseudo-)random float in the domain $[0, 1)$ for a floating-point format with a p-bit significand, many libraries simply divide a pseudo-random integer taken from $[0, 2^p - 1]$ by 2^p (Goualard, 2020). Does this ensure an even distribution of the bits in the fractional parts of the random floats, as required by applications such as *differential privacy* (Dwork, 2006; Mironov, 2012)? This can be systematically and quickly checked for a small floating-point format. We start by loading `MicroFloatingPoints` and `PyPlot` (alternatively, `PythonPlot` could also be used) for the graphics:

```
using MicroFloatingPoints, PyPlot
```

and we define a new IEEE 754-compliant floating-point type, say, with 7 bits for the exponent and 9 bits for the significand (i.e., 8 bits for the fractional part):

```
const E = 7 # Size of the exponent part
const f = 8 # Size of the fractional part
const p = f+1 # Size of the significand
const MuFP = Floatmu{E,f} # New IEEE 754 type
```

We now divide all integers in $[0, 2^p - 1]$ by 2^p to obtain a `MuFP` float, for which we record the value of each bit of its fractional part. An array `T` with `f` cells will accumulate the number of occurrences of a '1' over all floats produced (specifically, `T[i]` will contain the number of times the `(f-i)`-th bit of the fractional part was a '1' so far—with the zeroth bit being the rightmost one, as usual).

```
T = zeros(UInt32, f)
for v in 0:(2^p-1)
    d = MuFP(v)/2^p
    fpart = bitstring(d)[2+E:end] # Isolating the fractional part
    for j in 1:f
        global T[j] += Int(fpart[j] == '1')
    end
end
```

We now normalize to $[0, 1]$ the number of occurrences, and display the results with a bar plot (Figure 1).

```
nT = map(x -> x/2^p,T)
plt.bar(1:f,nT)
plt.xticks(1:f,reverse(map((x)->string(x-1),1:f)))
plt.yticks(0:0.1:1)
```

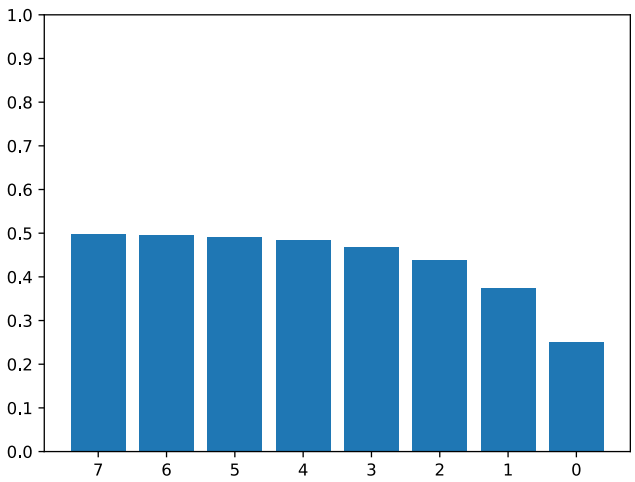

Figure 1: Probability of being '1' for each bit of the fractional part of a `Floatmu{7,8}` when dividing each integer in $[0, 2^9 - 1]$ by 2^9.

We were expecting a probability of 0.5 for each bit of the fractional part to be `1`. The actual plot shows that this is not the case and that the probability decreases for the lowest bits. It is very easy to check that behavior for a larger type by, e.g., changing the value of `f` to '16' in our previous code and reexecuting the script. The result in Figure 2 exhibits the same behavior for the larger type `Floatmu{7,16}`.

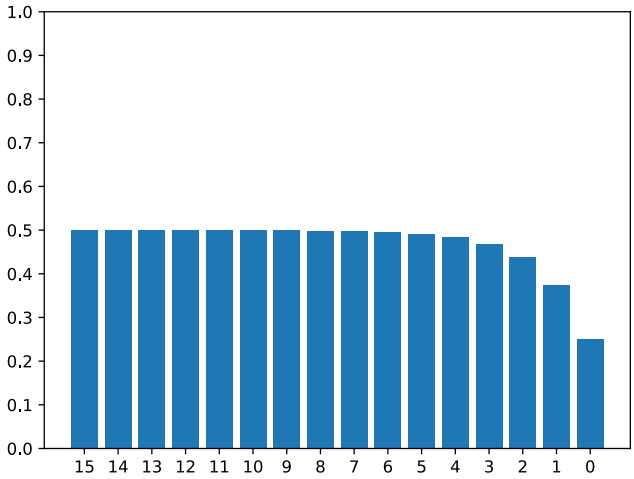

Figure 2: Probability of being '1' for each bit of the fractional part of a `Floatmu{7,16}` when dividing each integer in $[0, 2^{17} - 1]$ by 2^{17}.

Limitations

At present, all computations are performed in double precision (the `Float64` type), then correctly rounded to the `Floatmu{}` format chosen. As long as the precision of the `Floatmu{}` type is at most half the one of `Float64`, there is no *double rounding* issue (Martin-Dorel et al., 2013), and any final result obtained in that way is exactly the same as the one we would obtain by computing directly with the `Floatmu{}` precision (Rump, 2016).

Small floating-point formats are increasingly used in machine learning algorithms, where the precision and range are less important than the capability to store and manipulate as many values as possible. There are already some established formats implemented in hardware (e.g., IEEE 754 `Float16`, available natively in Julia, and Google Brain `bfloat16`, provided by the Julia Package `BFloat16s.jl`). There is, however, still a need for more flexibility to test the behavior of algorithms with varying precisions and ranges. The parametric type of `MicroFloatingPoints.jl` can be put to good use there too, and has already been for the study of training neural networks (Arthur et al., 2023). However, since it represents all floating-point formats by a pair of 32 bit integers, it cannot compete with more specialized packages for applications that require storing and manipulating massive amounts of numbers. For such use cases, it should therefore be confined to preliminary investigations with more limited amounts of data.

Acknowledgments

The reviewers and the editor for JOSS suggested many improvements to both the manuscript and the library itself during the reviewing process, which significantly increased their quality.

References

Arthur, B. J., Kim, C. M., Chen, S., Preibisch, S., & Darshan, R. (2023). A scalable implementation of the recursive least-squares algorithm for training spiking neural networks. *Frontiers in Neuroinformatics*, *17*. https://doi.org/10.3389/fninf.2023.1099510

Dwork, C. (2006). Differential privacy. In M. Bugliesi, B. Preneel, V. Sassone, & I. Wegener (Eds.), *Automata, languages and programming* (pp. 1–12). Springer. https://doi.org/10.1007/11787006_1

Fasi, M., & Mikaitis, M. (2023). CPFloat: A C library for simulating low-precision arithmetic. *ACM Transactions on Mathematical Software*, *49*(2), 18:1–18:32. https://doi.org/10.1145/3585515

Flegar, G., Scheidegger, F., Novaković, V., Mariani, G., Tomás, A. E., Malossi, A. C. I., & Quintana-Ortí, E. S. (2019). FloatX: A C++ library for customized floating-point arithmetic. *ACM Transactions on Mathematical Software*, *45*(4), 1–23. https://doi.org/10.1145/3368086

Goualard, F. (2020). Generating random floating-point numbers by dividing integers: A case study. In V. Krzhizhanovskaya (Ed.), *Proceedings of the international conference on computational science* (Vol. 12138, pp. 15–28). Springer. https://doi.org/10.1007/978-3-030-50417-5_2

Lefèvre, V. (2013). SIPE: Small integer plus exponent. *Proceedings of the 2013 IEEE 21st Symposium on Computer Arithmetic*, 99–106. https://doi.org/10.1109/ARITH.2013.22

Martin-Dorel, É., Melquiond, G., & Muller, J.-M. (2013). Some issues related to double rounding. *BIT Numerical Mathematics*, *53*(4), 897–924. https://doi.org/10.1007/s10543-013-0436-2

Mironov, I. (2012). On significance of the least significant bits for differential privacy. *Proceedings of the 2012 ACM Conference on Computer and Communications Security - CCS '12*, 650–661. https://doi.org/10.1145/2382196.2382264

Rump, S. M. (2016). IEEE754 precision-k base-β arithmetic inherited by precision-m base-β arithmetic for $k < m$. *ACM Transactions on Mathematical Software*, *43*(3), 20:1–20:15. https://doi.org/10.1145/2785965