Matrix multiplication over word-size modular rings using Bini's approximate formula *

B. Boyer[(a)], J.-G. Dumas[(b)]
[(a)] LIP6, UPMC, Paris, France
[(b)] LJK, Université de Grenoble, Grenoble, France
brice.boyer@lip6fr,jean.guillaume-dumas@imag.fr

A fast reliable matrix multiplication implementation over $\mathbf{Z}/p\mathbf{Z}$ is crucial in exact linear algebra. Indeed, many algorithms rely on fast matrix multiplication as a building block.

Bini's approximate formula (or border rank) for matrix multiplication [1] achieves a better complexity than Strassen's matrix multiplication formula [4]. We show a novel way to use the approximate formula in the special case where the ring is $\mathbf{Z}/p\mathbf{Z}$. Besides, we show an implementation à la FFLAS–FFPACK [3], where p is a word-size modulo, that improves on state-of-the-art $\mathbf{Z}/p\mathbf{Z}$ matrix multiplication implementations.

Bini's formula. Bini's approximate formula computes a matrix $C_\epsilon = A \times B + \epsilon D(\epsilon)$, with $A \in \mathbf{K}^{3\times 2}$, $B \in \mathbf{K}^{2\times 2}$ (noted $(3,2,2)$ multiplication), where D is a polynomial in $\mathbf{K}^{3\times 2}[\epsilon]$, and with 10 multiplications.

Application to exact matrix multiplication. We apply a different method than [1], requiring only one call to the approximate multiplication, for the special case $\mathbf{Z}/p\mathbf{Z}$. We are interested in the following two cases. First, we consider $\epsilon = 2^{-27}$ and use `double` floating point machine words; the idea is to store two exact integers in one `double` as $x + \epsilon y$, then any term in ϵ^2 will be neglected, as ϵ^2 approaches the machine precision (a rounding to the nearest will remove the $\epsilon-$approximations). Second, we take $\epsilon = p$, and the $p-$approximations are removed by a final reduction modulo p.

*This material is based on work supported in part by the National Science Foundation under Grant CCF-1115772 (Kaltofen) and Agence Nationale pour la Recherche under Grant ANR-11-BS02-013 HPAC (Dumas).

Proposition 1 (Case $\epsilon = 2^{-27}$) *For $\epsilon = 2^{-27}$ and a (m, k, n) matrix multiplication on $\mathbf{Z}/p\mathbf{Z}$, rounding to the nearest integer the output of one call to Bini $(3, 2, 2)-$approximate formula with* `double` *floating point arithmetic, gives the exact result when : $2\lfloor k/2 \rfloor (p-1)^2 < \frac{1}{3}2^{27}$.*

Proposition 2 (Case $\epsilon = p$) *For $\epsilon = p$ and a (m, k, n) matrix multiplication over $\mathbf{Z}/p\mathbf{Z}$, the reduction modulo p of the output C_ϵ of one call to Bini's $(3, 2, 2)-$approximate formula with* `double` *floating point arithmetic, gives the exact result when : $\lfloor k/2 \rfloor (p-1)^2(p+1)^2 < 2^{53}$.*

Remark. These bounds can be improved using a balanced representation.

Memory usage and scheduling We provide schedules requiring less extra memory (temporaries) than Strassen–Winograd's, in a similar fashion to [2], and implement them. We use only two temporaries and can create in-place algorithms by allowing overwriting an operand.

Implementation and Timings. Timings show that our implementation is competitive with Winograd's algorithm implementation, usually providing an $\approx 5\%$ speed-up, and it is always faster than FFLAS on `double`. The balanced representation allows to gain an $\approx 10\%$ speed-up on size 3 900 where the standard representation could not be used. The best speed-up of $\approx 15\%$ around sizes 2 700 to 3300 could be explained by optimal size BLAS block calls. For small moduli, the `float` representation performs better, but this phenomenon is only relevant for small moduli (≈ 400 and less, due to BLAS routines on `float` up to twice as fast as BLAS on `double`.

Bibliography

[1] D. Bini, *Relations between exact and approximate bilinear algorithms. Applications*, Calcolo 17 (1980), pp 87–97, Issue 1.

[2] B. Boyer, J.-G. Dumas, C. Pernet and W. Zhou *Memory efficient scheduling of Strassen-Winograd's matrix multiplication algorithm. Proc. of the 2009 ISSAC* (New York, NY, USA, 2009), ISSAC '09, ACM, pp. 55–62.

[3] J.-G. Dumas, P. Giorgi and C. Pernet *Dense linear algebra over word-size prime fields : the FFLAS and FFPACK packages.* ACM Trans. Math. Softw. 35, 3 (2008), 1–42.

[4] V. Strassen *Gaussian elimination is not optimal.* Numerische Mathematik 13 (1969), 354–356.