Fast and Small: Multiplying Polynomials without Extra Space

Daniel S. Roche

Symbolic Computation Group
School of Computer Science
University of Waterloo

CECM Day
SFU, Vancouver, 24 July 2009
Preliminaries

We study algorithms for univariate polynomial multiplication:

The Problem

<table>
<thead>
<tr>
<th>Given:</th>
<th>A ring ( R ), an integer ( n ), and ( f, g \in R[x] ) with degrees less than ( n )</th>
</tr>
</thead>
<tbody>
<tr>
<td>Compute:</td>
<td>Their product ( f \cdot g \in R[x] )</td>
</tr>
</tbody>
</table>

The Model

- Ring operations have unit cost
- Random reads from input, random reads/writes to output
- Space complexity determined by size of auxiliary storage
# Univariate Multiplication Algorithms

<table>
<thead>
<tr>
<th></th>
<th>Time Complexity</th>
<th>Space Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classical Method</strong></td>
<td>$O(n^2)$</td>
<td>$O(1)$</td>
</tr>
<tr>
<td><strong>Divide-and-Conquer</strong></td>
<td>$O(n^{\log_2 3})$ or $O(n^{1.59})$</td>
<td>$O(n)$</td>
</tr>
<tr>
<td>Karatsuba/Ofman ’63</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>FFT-based</strong></td>
<td>$O(n \log n \log \log n)$</td>
<td>$O(n)$</td>
</tr>
<tr>
<td>Schönhage/Strassen ’71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cantor/Kaltofen ’91</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>
## Univariate Multiplication Algorithms

<table>
<thead>
<tr>
<th></th>
<th>Time Complexity</th>
<th>Space Complexity</th>
</tr>
</thead>
<tbody>
<tr>
<td><strong>Classical Method</strong></td>
<td>$O(n^2)$</td>
<td>$O(1)$</td>
</tr>
<tr>
<td><strong>Divide-and-Conquer</strong></td>
<td>$O(n^{\log_2 3})$ or $O(n^{1.59})$</td>
<td>$O(n)$</td>
</tr>
<tr>
<td>Karatsuba/Ofman ’63</td>
<td></td>
<td></td>
</tr>
<tr>
<td><strong>FFT-based</strong></td>
<td>$O(n \log n \log \log n)$</td>
<td>$O(n)$</td>
</tr>
<tr>
<td>Schönhage/Strassen ’71</td>
<td></td>
<td></td>
</tr>
<tr>
<td>Cantor/Kaltofen ’91</td>
<td></td>
<td></td>
</tr>
</tbody>
</table>

Goal: Keep time complexity the same, reduce space
The Evolution of Multiplication

Small and slow
The Evolution of Multiplication

Big and fast
The Evolution of Multiplication

Small and fast
Previous Work

- **Savage & Swamy 1979**: $O(n^2)$ time-space lower bound for straight line programs.
- **Abrahamson 1985**: $O(n^2)$ time-space lower bound for branching programs.
Previous Work

- **Savage & Swamy 1979**: $O(n^2)$ time-space lower bound for straight line programs
- **Abrahamson 1985**: $O(n^2)$ time-space lower bound for branching programs
- **Monagan 1993**: Importance of space efficiency for multiplication over $\mathbb{Z}_p[x]$  
- **Maeder 1993**: Bounds extra space for Karatsuba multiplication so that storage can be preallocated — about $2n$ extra memory cells required.
- **Thomé 2002**: Karatsuba multiplication for polynomials using $n$ extra memory cells.
Present Contributions

- New Karatsuba-like algorithm with $O(\log n)$ space
- New FFT-based algorithm with $O(1)$ space under certain conditions
- Implementations in C over $\mathbb{Z}/p\mathbb{Z}$
Standard Karatsuba Algorithm

**Idea:** Reduce one degree-\(2k\) multiplication to three of degree \(k\).

- Originally noticed by Gauss (multiplying complex numbers), rediscovered and formalized by Karatsuba & Ofman

**Input:** \(f, g \in \mathbb{R}[x]\) each with degree less than \(2k\).

Write \(f = f_0 + f_1 x^k\) and \(g = g_0 + g_1 x^k\).
Low-Space Karatsuba Algorithms

Version “0”

Read-Only Input Space:

\[
\begin{array}{c}
\text{f01} \\
\text{f11}
\end{array}
\quad
\begin{array}{c}
g0 \\
g1
\end{array}
\]

Read/Write Output Space:

\[
\begin{array}{c}
\text{(empty)} \\
\text{(empty)} \\
\text{(empty)} \\
\text{(empty)}
\end{array}
\]

To Compute: \(f \cdot g\)
Low-Space Karatsuba Algorithms

Version “1”

1. The low-order coefficients of the output are initialized as $h$, and the product $f \cdot g$ is added to this.

Read-Only Input Space:

$$\begin{align*}
\text{f0} & \text{f1} \\
\text{g0} & \text{g1}
\end{align*}$$

Read/Write Output Space:

$$\begin{align*}
\text{h0} & \text{h1} \\
(\text{empty}) & (\text{empty})
\end{align*}$$

To Compute: $f \cdot g + h$
Low-Space Karatsuba Algorithms

Version “2”

1. The low-order coefficients of the output are initialized as \( h \), and the product \( f \cdot g \) is added to this.

2. The first polynomial \( f \) is given as a sum \( f^{(0)} + f^{(1)} \).

Read-Only Input Space:

\[
\begin{array}{c|c}
\text{f01} & \text{f11} \\
\hline
\text{f00} & \text{f10} \\
\end{array}
\quad
\begin{array}{c|c}
\text{g0} & \text{g1} \\
\end{array}
\]

Read/Write Output Space:

\[
\begin{array}{c|c}
\text{h0} & \text{h1} \\
\hline
(\text{empty}) & (\text{empty}) \\
\end{array}
\]

To Compute: \( (f^{(0)} + f^{(1)}) \cdot g + h \)
Dirty Details

Restrict modulus to 29 bits to allow for delayed reductions

In the Karatsuba step

- Only 4 values are added/subtracted in one position
- Delay reductions, perform two “corrections”

Classical algorithm

- Switch over at $n \leq 32$ (determined experimentally)
- Perform arithmetic in double-precision long longs; delay reductions (a la Monagan)
Problem: code explosion

3 “versions” of algorithms (based on extra constraints)
  ×
  Karatsuba or classical
  ×
  odd-sized or even-sized operands
  ×
  equal-sized operands or “one different”

Solution: Use “supermacros” in C:
Same file is included multiple times with some parameter values changed (crude form of code generation).
DFT-Based Multiplication

**Input**
- $f$
- $g$

**Evaluation (DFT)**
- $\text{DFT}(f)$
- $\text{DFT}(g)$

**Pointwise multiplication**
- $\ast$

**Interpolation (inverse DFT)**
- $f \cdot g$
- $\text{DFT}(f \cdot g)$
Simplifying Assumptions

From now on:

- $\deg f + \deg g < n = 2^k$ for some $k \in \mathbb{N}$
- The base ring $R$ contains a $2^k$-PRU $\omega$

That is, assume “virtual roots of unity” have already been found, and optimize from there.
Usual Formulation of the FFT

Perform two $\frac{n}{2}$-DFTs followed by $\frac{n}{2}$ 2-DFTs:

- Write $f(x) = f_{\text{even}}(x^2) + x \cdot f_{\text{odd}}(x^2)$  
  (i.e. $\deg f_{\text{even}}, \deg f_{\text{odd}} < n/2$)
- Compute $\text{DFT}_{\omega^2}(f_{\text{even}})$ and $\text{DFT}_{\omega^2}(f_{\text{odd}})$
- Compute each $f(\omega^i) = f_{\text{even}}(\omega^{2i}) + \omega \cdot f_{\text{odd}}(\omega^{2i})$

Make use of “butterfly circuit” for each size-2 DFT:

\[
\begin{align*}
    a & \quad \leftrightarrow \quad a + \omega^i \cdot b \\
    b & \quad \leftrightarrow \quad a - \omega^i \cdot b
\end{align*}
\]
Example: 8-Way FFT

The diagram illustrates an 8-way FFT with inputs $a_0, a_1, a_2, a_3, a_4, a_5, a_6, a_7$ and outputs $f(\omega^0), f(\omega^1), f(\omega^2), f(\omega^3), f(\omega^4), f(\omega^5), f(\omega^6), f(\omega^7)$.
Reverted Binary Ordering

In-Place FFT permutes the ordering into reversed binary:

\begin{align*}
0 &= 000_2 \\
1 &= 001_2 \\
2 &= 010_2 \\
3 &= 011_2 \\
4 &= 100_2 \\
5 &= 101_2 \\
6 &= 110_2 \\
7 &= 111_2
\end{align*}

Problem: Powers of $\omega$ are not accessed in order
Possible solutions:

- Precompute all powers of $\omega$ — too much space
- Perform steps out of order — terrible for cache
- Permute input before computing — costly
Alternate Formulation of FFT

Perform \( \frac{n}{2} \) 2-DFTs followed by two \( \frac{n}{2} \)-DFTs

- Write \( f = f_{\text{low}} + x^{n/2} \cdot f_{\text{high}} \)
- Compute \( f_0 = f_{\text{low}} + f_{\text{high}} \) and \( f_1 = f_{\text{low}}(\omega x) - f_{\text{high}}(\omega x) \)
- Compute each \( f(\omega^2i) = f_0(\omega^2i) \) and \( f(\omega^{2i+1}) = f_1(\omega^2i) \)

Modified “butterfly circuit”:

\[ a \rightarrow i \rightarrow a + b \]
\[ b \rightarrow i \rightarrow (a - b)\omega^i \]
Example: 8-Way In-Place FFT (Alternate Formulation)
Folded Polynomials

Recall the basis for the “alternate” FFT formulation:

\[ f_0 = f_{\text{low}} + f_{\text{high}} \]
\[ f_1 = f_{\text{low}}(\omega x) - f_{\text{high}}(\omega x) \]

A generalization (recalling that \( n = 2^k \)):

**Definition (Folded Polynomials)**

\[ f_i = f(\omega^{2^{i-1}} x) \mod x^{2^{k-i}} - 1 \]

**Theorem**

\[ f(\omega^{2^i(2j+1)}) = f_{i+1}(\omega^{2^{i+1}j}) \]

So by computing each \( f_i \) at all powers of \( \omega^i \), we get the values of \( f \) at all powers of \( \omega \).
Recursively Applying the Alternate Formulation

Example (Reverted Binary Ordering of 0, 1, …, 15)

\[0, 8, 4, 12, 2, 10, 6, 14, 1, 9, 5, 13, 3, 11, 7, 15\]

\(\text{DFT}_\omega(f)\) in binary reversed order can be computed by DFTs of \(f_i\)s:
FFT-Based Multiplication without Extra Space

**Idea:** Solve half of remaining problem at each iteration

\[ f \quad \text{Input} \quad g \]

(empty)
FFT-Based Multiplication without Extra Space

Idea: Solve half of remaining problem at each iteration
**Idea**: Solve half of remaining problem at each iteration

In-Place FFTs (alternate formulation)
**Idea:** Solve half of remaining problem at each iteration

Pointwise Multiplication

- **DFT**($f_1$)  
- **DFT**($f_1 \cdot g_1$)

Illustration:
- Function $f$ represented by a green block.
- Function $g$ represented by a blue block.
- DFT($f_1$) and DFT($f_1 \cdot g_1$) are represented by red and orange blocks, respectively.
- The asterisk (*) symbolizes pointwise multiplication.
FFT-Based Multiplication without Extra Space

Idea: Solve half of remaining problem at each iteration

Foldiing
FFT-Based Multiplication without Extra Space

**Idea:** Solve half of remaining problem at each iteration

<table>
<thead>
<tr>
<th>f</th>
<th>g</th>
</tr>
</thead>
</table>

**In-Place FFTs (alternate formulation)**

```
DFT(f2)  DFT(g2)  DFT(f1·g1)
```
FFT-Based Multiplication without Extra Space

**Idea**: Solve half of remaining problem at each iteration

\[
\text{Pointwise Multiplication}
\]

\[
\begin{align*}
\text{DFT}(f) & \quad \text{DFT}(f_2 \cdot g_2) & \quad \text{DFT}(f_1 \cdot g_1)
\end{align*}
\]
FFT-Based Multiplication without Extra Space

**Idea:** Solve half of remaining problem at each iteration

(f · g)
FFT-Based Multiplication without Extra Space

**Idea**: Solve half of remaining problem at each iteration

\[ f \cdot g \]
Analysis

Time cost of the various stages:

- **Folding**: $O(n)$ cost times $\log n$ folds = $O(n \log n)$
- **FFTs**: $O(m \log m)$ for $m = n, n/2, n/4, \ldots, 1 = O(n \log n)$
- **Multiplications**: $n/2 + n/4 + \cdots + 1 = O(n)$

Total cost: $O(n \log n)$ time and $O(1)$ extra space when the following conditions hold:

- $n = \deg f + \deg g + 1$ is a power of 2
- $R$ contains an $n$-PRU $\omega$
Modular Arithmetic

Use floating-point Barrett reduction (from NTL):

- Pre-compute an approximation of $1/p$
- Given $a, b \in \mathbb{Z}_p$, compute an approximation of $q = \lfloor a \cdot b \cdot (1/p) \rfloor$
- Then $ab - qp$ equals $ab \text{rem} p$ plus or minus $p$.

The cost of this method:

- 2 double multiplications
- 2 int multiplications
- 1 int subtraction
- 3 conversions between int and double
- 2 “correction” steps to get exact result

↔ not necessary until the very end!
Implementation Benchmarking

Details of tests:

- 2.5 GHz 64-bit Athalon, 256KB L1, 1MB L2, 2GB RAM
- \( p = 167772161 \) (28 bits)
- Comparing CPU time (in seconds) for the computation

Disclaimer

We are comparing apples to oranges.
Timing Benchmarks

The graph compares the time taken for operations over NTL with different algorithms. The x-axis represents $\log_2 n$, and the y-axis shows the time taken over NTL. The graph depicts two types of algorithms: Karatsuba-like and FFT-based.
Future Directions

- Efficient implementation over $\mathbb{Z}$ (GMP)
- Similar results for Toom-Cook 3-way or $k$-way
- What modulus bit restriction is “best”?
- Is completely in-place (overwriting input) possible?