Interval Partitions and Polynomial Factorization

Daniel Panario
School of Mathematics and Statistics
Carleton University
daniel@math.carleton.ca

Joint work with
J. von zur Gathen and B. Richmond

Fq9, July 2009

The problem

Let $\mathbb{F}_q$ be a finite field with q elements:

Given a monic univariate polynomial $f \in \mathbb{F}_q[x]$, find the complete factorization $f = f_1^{e_1} \cdots f_r^{e_r}$, where the f_i's are monic distinct irreducible polynomials and $e_i > 0$, $1 \leq i \leq r$.

Applications

- Algebraic coding theory (Berlekamp 1968);
- Computer algebra (Collins 1979, Knuth 1981, Geddes, Czapor and Labahn 1992);
- Cryptography (Chor and Rivest 1984, Odlyzko 1985, Lenstra 1991);
- Computational number theory (Buchmann 1990).

A general factoring method

A basic factorization algorithm

ERF *Elimination of repeated factors* replaces a polynomial by a squarefree one which contains all the irreducible factors of the original polynomial with exponents reduced to 1.

DDF *Distinct-degree factorization* splits a squarefree polynomial into a product of polynomials whose irreducible factors have all the same degree.

EDF *Equal-degree factorization* factors a polynomial whose irreducible factors have the same degree.

The first step in the factorization chain of a polynomial is the *elimination of repeated factors* (ERF). It essentially accounts for a gcd between the polynomial to be factored and its derivative.

This method has similar cost to the squarefree factorization methods. Its cost is negligible when compared with the other steps of the algorithm.

The second step *distinct-degree factorization* (DDF) is based on the following theorem.

Theorem. *For $i \geq 1$, the polynomial $x^{q^i} - x \in \mathbb{F}_q[x]$ is the product of all monic irreducible polynomials in $\mathbb{F}_q[x]$ whose degree divides i.*

The third step *equal-degree factorization* (EDF) involves factoring polynomials b_k that have all their irreducible factors of the same (known) degree k. The reference is Cantor-Zassenhaus' probabilistic algorithm.

The Chinese remainder theorem implies

$$\mathbb{F}_q[x]/(b) \cong \mathbb{F}_q[x]/(f_1) \times \cdots \times \mathbb{F}_q[x]/(f_j).$$

The test $h_i^{(q^k-1)/2} = 1$ discriminates the squares in the multiplicative group of $\mathbb{F}_q[x]/(f_i)$. Taking a random h and computing $a := h^{(q^k-1)/2} - 1 \mod b$, we have that $\gcd(a, b)$ "extracts" the product of all the f_i for which h is a square in $\mathbb{F}_q[x]/(f_i)$.

EDF can be done faster than DDF using a randomized method.

Many authors indicate that the most time-consuming part of the algorithm is the distinct-degree factorization.

Bottleneck of the method: DDF.

Let's assume that we have no knowledge of the polynomial being factored. Then, it is natural to assume that the polynomial is taken uniformly at random.

Theorem. (Flajolet, Gourdon and Panario, 2001)
(i) The probability that DDF yields the complete factorization is asymptotic to

$$c_q = \prod_{k\geq 1}\left(1+\frac{I_k}{q^k-1}\right)(1-q^{-k})^{I_k},$$

$$c_2 \doteq 0.6656,\ c_{257} \doteq 0.5618,\ c_\infty = e^{-\gamma} \doteq 0.5614.$$

(ii) The number of degree values for which there is more than one irreducible factor in the polynomial produced by DDF has an average that is asymptotic to the constant

$$\sum_{k\geq 1}(1-q^{-k})^{I_k}\left((1-q^{-k})^{-I_k}-1-\frac{I_k q^{-k}}{1-q^{-k}}\right).$$

(iii) The degree of the part of the polynomial that remains to be factored by the EDF algorithm has expectation $\log n + O(1)$*, and standard deviation of approximately* $\sqrt{n}$*.*

One drawback of the algorithm is that most of the gcds computed will be equal to 1, since a random polynomial of degree n has about $\log n$ irreducible factors on average.

How can we save gcd computations?

Interval partition

To reduce the number of gcd computations, von zur Gathen and Shoup (1992) and Kaltofen and Shoup (1995) present algorithms for the DDF step based on a baby-step giant-step strategy:

> *Divide the interval* $1, \ldots, n$ *into about* $\sqrt{n}$ *intervals of size* $\sqrt{n}$*; for each interval, compute the joint product of the irreducible factors whose degree lies in that interval. Use DDF for every interval with more than one irreducible factor.*

An interval partition of $[1 \ldots n]$ is a sequence $S = (s_0, \ldots, s_m)$ of integers with $0 = s_0 < s_1 < \cdots < s_m = n$. The intervals of the partition are the sets $\pi_j = \{s_{j-1} + 1, \ldots, s_j\}$ for $1 \leq j \leq m$.

A coarse DDF computes a *partial factorization* $f = f_1 \cdot f_2 \cdots$ where f_j is the product of all irreducible factors of the original polynomial with degrees belonging to π_j.

If f_j contains at most one irreducible factor, there is no need of further computation. Otherwise, a fine DDF is executed for this partial factorization using DDF.

An interval polynomial for an interval $\pi_j = \{s_{j-1} + 1, \ldots, s_j\}$ is a polynomial that is divisible by any irreducible factor whose degree lies in π_j.

Interval polynomials:

• von zur Gathen and Shoup (1992): $\prod_{i \in \pi_j} x^{q^i} - x$ is divisible by every irreducible polynomial in $\mathbb{F}_q[x]$ of degree dividing any $i \in [s_{j-1} + 1, s_j]$.

• Kaltofen and Shoup (1995) and Shoup (1995): $\prod_{0 \leq i \leq s_j - s_{j-1}} x^{q^{s_j}} - x^{q^i}$ based on the following theorem (Kaltofen and Shoup, 1995).

Theorem. For nonnegative integers $i > j$, the polynomial $x^{q^i} - x^{q^j} \in \mathbb{F}_q[x]$ is divisible by those irreducible polynomials in $\mathbb{F}_q[x]$ whose degree divides $i - j$.

The algorithms by von zur Gathen and Shoup (1992) and Kaltofen and Shoup (1995) split the interval $[1 \ldots n]$ into about $\sqrt{n}$ pieces of size $\sqrt{n}$ each. When dealing with random polynomials, this breaking strategy is not the best possible.

The number of irreducible factors in a random polynomial of degree n tends to a Gaussian distribution with mean value $\log n$.

These $\log n$ factors are not equally distributed in the interval $[1, n]$: the expected number of irreducible factors of degree k in a random polynomial is roughly $1/k$. Thus, one expects to have more factors of lower degrees than of higher degrees.

When dealing with random polynomials, it is natural to consider partitions with growing interval sizes in order to avoid collision of irreducible factors in intervals.

von zur Gathen and Gerhard (2002) use polynomially growing interval sizes to factor large degree random polynomials over $\mathbb{F}_2$.

These intervals have led to the million-degree factorization of Bonorden, von zur Gathen, Gerhard, Müller and Nöcker (2000).

The analysis of these algorithms involve studying the degree distribution of irreducible factors in intervals (this work).

Results

We provide useful information on the parameters related to partitions of the interval $[1, n]$:

- mean value and variance for the number of multi-factor intervals of a polynomial (intervals with more than one irreducible factor);
- mean value and variance for the number of irreducible factors of a polynomial whose degrees lie in any of its multi-factor intervals;
- mean value and variance for the total degree of irreducible factors (of a polynomial) whose degrees lie in any of the multi-factor intervals for the polynomial;
- mean value and variance for the number of gcds executed;

and so on.

Number of gcds

The number of gcds executed is the addition of the number of gcds at the coarse DDF level (that is, the number m of parts of the interval partition) and the number of gcds at the fine DDF level.

For partitions of the form $s_k = k^j$, the number of gcds at the coarse level is roughly $n^{1/j}$.

For the number of gcds at the fine DDF level we assume that when an interval is multi-factor the number of gcds executed equals to the length of the interval (there is a faster algorithm that would stop as soon as we reach the second largest degree irreducible factor inside the multi-factor interval; see Flajolet, Gourdon and Panario, 2001).

Theorem. Let $j > 1$ be a real number, $s_k = k^j$ an interval partition of $[1 \ldots n]$ with intervals $\pi_1, \pi_2, \ldots$, and $d_k = s_k - s_{k-1}$. Then, the expected number of gcds executed in multi-factor intervals of a polynomial behaves, for $n \to \infty$, as follows:

- ◇ it converges to a constant for $j < 2$;
- ◇ it is asymptotic to $4(1 - 1/q) \ln n$ for $j = 2$; and
- ◇ it is asymptotic, for $j > 2$, to

$$\left(1 - \frac{1}{q}\right) \frac{j^3}{j-2} \frac{1}{2^{1-\frac{2}{j}}} n^{1-\frac{2}{j}}.$$

Proof (sketch).

The generating function marking the size d_k of the kth interval π_k, if it contains more than one irreducible factor, is

$$S_1(z,u) = \prod_{k\geq 1}\left(u^{d_k}\prod_{\ell\in\pi_k}\left(1+z^\ell\right)^{I_\ell} + (1-u^{d_k})\left(1+\sum_{\ell\in\pi_k} I_\ell z^\ell\right)\right),$$

where we consider squarefree polynomials since we are in the distinct-degree stage. The coefficient $[z^n u^i]S_1(z,u)$ equals the number of squarefree polynomials of degree n that require i gcds in multi-factor intervals of the given partition.

The mean value of the number of gcds in multi-factor intervals for a polynomial is obtained by differentiating $S_1(z,u)$ with respect to u, and then setting $u=1$; we get

$$\left.\frac{\partial S_1(z,u)}{\partial u}\right|_{u=1} = \frac{1}{1-qz} Q_1(z),$$

where

$$Q_1(z) = (1-qz^2)\left(\sum_{k\geq 1} d_k \left(1 - \prod_{\ell\in\pi_k}(1+z^\ell)^{-I_\ell}\left(1+\sum_{\ell\in\pi_k} I_\ell z^\ell\right)\right)\right).$$

Using the standard expression for I_ℓ and the change $z=t/q$, we obtain the approximation

$$1 - \prod_{\ell\in\pi_k}\left(1+z^\ell\right)^{-I_\ell}\left(1+\sum_{\ell\in\pi_k} I_\ell z^\ell\right) \sim \left(t^{s_{k-1}+1} d_k/(s_{k-1}+1)\right)^2.$$

Consider partitions of $[1 \ldots n]$ of the form $s_k = k^j$. We have $s_{k-1} = (k-1)^j$ and $d_k = s_k - s_{k-1} \sim jk^{j-1}$:

$$\begin{aligned} Q_1\left(\frac{t}{q}\right) &\sim \sum_{k\geq 1}\left(1-\frac{t^2}{q}\right) t^{2s_{k-1}+2} d_k^3/s_k^2 \\ &\sim \sum_{k\geq 1}\left(1-\frac{t^2}{q}\right) t^{2(k-1)^j+2} j^3 k^{j-3}. \end{aligned} \tag{1}$$

We immediately conclude that for $n \to \infty$ and $j < 2$, the expected number of gcds executed in multi-factor intervals of a polynomial converges to a constant.

The case $j > 2$ and $j = 2$ can be treated in a similar (but slightly more complicated) way. Technically, they require Flajolet and Oldlyzko singularity analysis.

■

To compute the variance, using similar techniques as before, we first compute the second moment by differentiating $S_1(z,u)$ with respect to u two times and putting $u=1$. We obtain, using singularity analysis, that the second moment is asymptotic to

$$\left(1-\frac{1}{q}\right)\frac{j^3}{2^{2-3/j}}\frac{\Gamma(2-3/j)}{\Gamma(3-3/j)}n^{2-3/j}=\left(1-\frac{1}{q}\right)\frac{j^4}{2^{2-3/j}(2j-3)}n^{2-3/j}.$$

Since the order of the expected value for the number of gcds executed at the fine DDF level is constant, $\log n$ or $n^{1-2/j}$, the variance is given by the second moment. We have a standard deviation of order $n^{1-3/(2j)}$.

Theorem. The variance of the number of gcds executed in the factorization process has asymptotic order $n^{2-3/j}$.

Conclusions

For partitions of the form $s_k = k^j$, for $j > 1$ we get

- For $1 < j < 2$, the total number of gcds is governed by the coarse DDF level at a cost of roughly $n^{1/j}$ gcds.
- For $j = 2$, the gcds at the fine DDF level show some weight $(4 \ln n)$, but overall the number of gcds is determined by the coarse level at a cost of $\sqrt{n}$ gcds.
- For $j > 2$, we have $n^{1/j}$ gcds at the coarse DDF level and $\left(1 - \frac{1}{q}\right) \frac{j^3}{j-2} \frac{1}{2^{1-\frac{2}{j}}} n^{1-\frac{2}{j}}$ gcds at the fine DDF level. We get that in the range $2 < j < 3$ the cost is governed by the coarse DDF level, while in the range $j > 3$ the cost is determined by the fine DDF algorithm. At $j = 3$, both exponent are the same, giving order $n^{1/3}$ gcds for the whole process.

The best partition of the form $s_k = k^j$, for $j > 1$, in terms of minimizing the upper bound on the number of gcds is $s_k = k^3$.

Further work

In this work we establish the first steps towards a full analysis of interval parameters for polynomial factorization over finite fields.

Future work includes:

- Estimation of the number of the gcds inside a multi-factor interval stopping when the largest degree irreducible factor is processed.
- Analysis of other partitions different from $s_k = k^j$, for $j > 1$.
- Actual estimation of the cost of the algorithms in terms of operations over $\mathbb{F}_q$.