Random Polynomials and Polynomial Factorization

Philippe Flajolet, Xavier Gourdon, Daniel Panario

INRIA

INSTITUT NATIONAL DE RECHERCHE EN INFORMATIQUE ET AUTOMATIQUE

Random Polynomials and Polynomial Factorization

Philippe FLAJOLET, Xavier GOURDON, Daniel PANARIO

N° 2852
Mars 1996

PROGRAMME 2

ISSN 0249-6399

Random Polynomials and Polynomial Factorization

Philippe Flajolet, Xavier Gourdon and Daniel Panario

Abstract

We give a precise average-case analysis of a complete polynomial factorization chain over finite fields by methods based on generating functions and singularity analysis.

Polynômes aléatoires et factorisation de polynômes

Résumé

Nous donnons une analyse en moyenne précise d'une chaîne complète de factorisation de polynômes sur les corps finis par des méthodes fondées sur les fonctions génératrices et l'analyse de singularités.

To appear in *Automata, Languages and Programming, Proceedings of the 23rd ICALP colloquium*, Paderborn, July 1996, F. Meyer auf der Heide, Ed., in *Lecture Notes in Computer Science*.

Random Polynomials and Polynomial Factorization

Philippe Flajolet,[1] Xavier Gourdon,[1] and Daniel Panario[2]

[1] Algorithms Project, INRIA Rocquencourt, F-78153 Le Chesnay, France.
[2] Department of Computer Science, University of Toronto, Toronto, Canada M5S-1A4.
E-mails: Philippe.Flajolet@inria.fr, Xavier.Gourdon@inria.fr, daniel@cs.toronto.edu.

Abstract. We give a precise average-case analysis of a complete polynomial factorization chain over finite fields by methods based on generating functions and singularity analysis.

1. Introduction

Polynomial factorization is basic to many areas of computer algebra [12], algebraic coding theory [1], computational number theory and cryptography [2, 6, 18, 20]. Its implications include finding complete partial fraction decompositions (a problem itself useful for symbolic integration), designing cyclic redundancy codes, computing the number of points on elliptic curves and building arithmetic public key cryptosystems.

Polynomial factorization may be carried out over any field, but the efficient algorithms are essentially *probabilistic* and they eventually rely on factoring over a *finite field* $\mathbb{F}_q$ where q is a prime or the power of a prime, see [16] for an excellent introduction. This paper derives basic properties of random polynomials over finite fields that are of interest in the study of factoring algorithms. We show that the most important characteristics can be treated systematically by methods of "analytic combinatorics" based on generating functions and singularity analysis.

We have elected here to consider a classical factorization chain over finite fields that is at the same time simple, fairly efficient, and *complete*. It is close to what is used internally in the Maple computer algebra system [12] and to what is likely to be required of a general purpose computer algebra system that mostly deals with polynomials of intermediate "size". Our factorization chain may not be the fastest at the moment, compare for instance with Shoup's technique [24]. However the discipline of completely analyzing such algorithms, which is in the line of Knuth's works [16], reveals parameters that are of intrinsic interest for polynomial factoring in general. To the best of our knowledge, such a task has not been undertaken systematically beyond rough (mostly worst-case) bounds.

Our reference factorization chain comprises the following three classical steps:

ERF: *Elimination of repeated factors* replaces a polynomial by a squarefree form that 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 all have the same degree.

EDF: *Equal-degree factorization* factors a polynomial the irreducible factors of which all have the same degree.

The top-level code of our factorization chain (in pseudo-Maple) is given below.

```
procedure factor(f : polynomial);
     1: a := ERF(f);
     2: b := DDF(a);
        F := 1;
     3: for k from 1 to n do
             F := F . EDF(b[k],k);
        od;
     4: return(F . factor(f/a));
end;
```

Computational model. All average-case analyses are expressed as asymptotic forms in n, the degree of the polynomial to be factored. We *fix a finite field* $\mathbb{F}_q$ with $q = p^m$ (p prime) and consider the polynomial ring $\mathbb{F}_q[x]$, see [12, 16, 19]. For simplicity of exposition, we assume here that the characteristic p is odd, but the algorithms and their analyses can be easily adapted to the otherwise important cases of $\mathbb{F}_2$ and $\mathbb{F}_{2^m}$. Our model assumes that a basic field operation has cost $\mathcal{O}(1)$; then the cost of a sum is $\mathcal{O}(n)$ and the cost of a product, a division or a gcd is $\mathcal{O}(n^2)$, when applied to polynomials of degree $\leq n$. For *dominant asymptotics*, we can freely restrict attention to polynomial products and gcd's whose costs can be taken under the standard form

$$\text{product: } \tau_1 n^2, \qquad \text{gcd: } \tau_2 n^2.$$

2. Summary of results

It is well-known [1, 16] that a random polynomial of degree n is irreducible with probability tending to 0 and has close to $\log n$ factors on average and with a high probability [4, 10]. Thus, the factorization of a random polynomial over a finite field is almost surely nontrivial.

The first phase *ERF* of our factorization chain classically starts with the elimination of repeated factors, a simplified form of squarefree factorization described in Section 4. Theorem 1 quantifies this process and shows that up to smaller order terms, the expected cost is dominated by a single gcd of the polynomial f to be factored and its derivative f', so that it is $\mathcal{O}(n^2)$ on average. In a precise technical sense, most of the factorization cost results from the subsequent phases since the non-squarefree part has average degree $\mathcal{O}(1)$.

The second phase *DDF* that is described in Section 5 splits the squarefree part a of the polynomial to be factored into a product $a = b_1 \cdot b_2 \cdots b_n$, where b_k is itself the product of the irreducible factors of a that have degree k. This phase is based on elementary properties of finite fields and is the one with the highest computational cost, namely $\mathcal{O}(n^3)$ on average. Theorems 3,4,5 provide a precise comparison of three strategies: the naïve rule, the "half-degree" rule and the "early abort" rule whose costs are found to be in the approximate proportion $1 :: \frac{3}{4} :: \frac{2}{3}$. Thus a savings of about one third results from controlling the DDF phase by the early abort strategy. At the end of this phase, the factorization is complete with a probability ranging asymptotically between 0.56 and 0.67, see Theorem 6.

The third phase *EDF* can be exactly analysed and it is found that its expected cost is comparatively small, being $\mathcal{O}(n^2)$, see Theorems 7,8 for precise statements. For each nontrivial factor b_k, it involves a recursive refinement process again based on properties of finite fields. The analysis is close to that of digital trees known as "tries" [15] but under a biased probability model.

Precise statements are given in the next few pages with an explicit dependency on the field cardinality q, and some of them involve number-theoretic functions that can be both evaluated and estimated easily. Therefore, the results obtained allow us to quantify precisely what goes on. A simplified picture is as follows. The ERF phase involves with high probability little more than a single polynomial gcd. The DDF phase of cost $\mathcal{O}(n^3)$ is the one that is most intensive computationally, where control by the "early-abort" strategy is expected to bring gains close to 36% at no extra cost. The last phase of EDF is executed less than 50% of the time and its cost is again small compared to that of DDF.

3. Basic methodology

This paper relies heavily on a symbolic use of *generating functions* (GF's). These are used to express enumerative properties of random polynomials and also to derive direct asymptotic results from singularities. General references are Chapter 3 of Berlekamp's book [1], the exercise section 4.6.2 of Knuth's book [16], and the paper by Flajolet and Odlyzko [9] for asymptotic methods.

3.1. Generating functions. We specialize our discussion to polynomials over a finite field $\mathbb{F}_q$. Let $\mathcal{I}$ be the collection of monic irreducible polynomials. The two expressions

$$\mathcal{Q} = \prod_{\omega \in \mathcal{I}} (1+\omega), \qquad \text{and} \qquad \mathcal{P} = \prod_{\omega \in \mathcal{I}} (1-\omega)^{-1}. \tag{1}$$

when expanded by distributivity "generate" formally the family $\mathcal{Q}$ of monic squarefree polynomials and $\mathcal{P}$ of all monic polynomials. In this context, $\mathcal{I}$ may itself be identified with the formal sum $\mathcal{I} = \sum_{\omega \in \mathcal{I}} \omega$.

Let z be a formal variable. The substitution $\omega \mapsto z^{|\omega|}$ with $|\omega|$ the degree of $\omega \in \mathcal{I}$ produces generating functions by a well-known process For instance, $I(z) = \sum_{\omega \in \mathcal{I}} z^{|\omega|} = \sum_n I_n z^n$, where I_n is the number of polynomials in $\mathcal{I}$ having degree n. The same substitution applied to $\mathcal{P}$ and $\mathcal{Q}$ yields two series, $P(z)$ and $Q(z)$, that are found to satisfy

$$Q(z) = \prod_{n=1}^{\infty} (1+z^n)^{I_n}, \qquad P(z) = \prod_{n=1}^{\infty} (1-z^n)^{-I_n}. \tag{2}$$

Then, the coefficients $Q_n = [z^n]Q(z)$ and $P_n = [z^n]P(z)$ represent the number of polynomials of degree n in $\mathcal{Q}$ and $\mathcal{P}$ respectively.

Since P_n has value q^n, we have $P(z) = (1-qz)^{-1}$, and the second relation of (2) implicitly determines I_n by a well-known process based on Moebius

inversion [1]

$$I_n = \frac{1}{n}\sum_{k|n}\mu(k)q^{n/k}, \qquad \text{so that} \qquad I_n = \frac{q^n}{n} + \mathcal{O}\left(\frac{q^{n/2}}{n}\right), \tag{3}$$

$$I(z) = \sum_{k=1}^{\infty}\frac{\mu(k)}{k}\log\frac{1}{1-qz^k}. \tag{4}$$

Thus a fraction extremely close to $1/n$ of all polynomials of degree n are irreducible. This result was first proven by Gauss [11] for prime fields (see also [8]).

As regards Q_n, the formula $1+z=(1-z^2)/(1-z)$ applied to the infinite products for $P(z), Q(z)$ entails

$$Q(z) = \frac{P(z)}{P(z^2)} = \frac{1-qz^2}{1-qz}, \qquad \text{and} \qquad Q_n = q^{n-1}(q-1) \quad (n\geq 2), \tag{5}$$

with $Q_0=1$, $Q_1=q$. Apparently, this result was given for the first time in [5].

Parameters. We need extensions of this symbolic method in order to take care of characteristic parameters of polynomial factorization. Let Φ be a class of monic polynomials, χ some integer–valued parameter on Φ. The sum

$$\Phi(z,u) = \sum_{\omega\in\Phi} z^{|\omega|}u^{\chi(\omega)}$$

is such that the coefficient $[z^n u^k]\Phi(z,u)$ represents the number of polynomials of degree n and χ–parameter equal to k. For additive parameters χ, the product decompositions above generalize, provided one uses the translation rule $\omega \mapsto z^{|\omega|}u^{\chi(\omega)}$. The technique of rearranging logarithms of infinite products is useful in simplifying such expressions.

Averages and standard deviations are obtained by taking successive derivatives of bivariate generating functions with respect to u, then setting $u=1$.

3.2. Asymptotic analysis. Generating functions (GFs) encode exact informations on their coefficients. Furthermore, their behaviour near their dominant positive singularity is an important source of coefficient asymptotics.

Most of the generating functions $f(z)$ to be studied in this paper are singular at $z=1/q$ with an isolated singularity of the algebraic-logarithmic type. In that case, an expansion near $z=1/q$ of the form

$$f(z) = \frac{1}{(1-qz)^{\alpha}}\left(\log\frac{1}{1-qz}\right)^k (1+o(1)). \tag{6}$$

is translated to coefficients by the method known as singularity analysis [9, 21]

$$[z^n]f(z) = q^n\frac{n^{\alpha-1}}{\Gamma(\alpha)}(\log n)^k\,(1+o(1)), \tag{7}$$

whenever $\alpha\neq 0,-1,-2,\ldots$. This requires analytic continuation (isolated singularity), a condition for instance satisfied by the GF's of Theorems 1,3,4.

The same translation can be effected under a variety of alternative conditions corresponding either to Darboux's method [7] or to Tauberian theorems of the

Hardy-Littlewood-Karamata type [13, 21, 22]. Such alternative conditions are needed in Theorems 5 and 6 where the GF's have a natural boundary.

The permutation model. The following property is well-known. The joint distribution of degrees in the prime decomposition of a random polynomial over $\mathbb{F}_q$ having degree n admits as a limit, when the cardinality q of the base field tends to infinity (n staying fixed!), the joint distribution of cycle lengths in random permutations of size n. Accordingly GF's of random polynomials at z/q converge to GF's of corresponding permutation families when $q \to +\infty$.

This gives rise to a useful heuristic for large field cardinalities. An instance is mentioned in [13] in connection with the probability that a random polynomial admits factors of distinct degrees which, for large q and large n is found to approach $e^{-\gamma}$. Our Theorem 6 illustrates an instance of this situation.

4. Elimination of repeated factors (ERF)

The first step in the factorization chain of a polynomial is the *elimination of repeated factors* (ERF). In characteristic 0, this is achieved by the gcd of f and its derivative f'. In finite characteristics, additional control is needed in order to deal with pth powers whose derivatives are 0, see [12, 16]. The auxiliary computation of pth roots, $g^{1/p}$, is performed in the classical way described in [12, p. 344] for example.

```
procedure ERF(f : polynomial);
        g := gcd(f,f'); h := f/g; k := gcd(g,h);
        while k<>1 do  g := g/k; k := gcd(g,h)  od;
        if  g <> 1 then h := h*ERF(g^(1/p)) fi;
        return(h);
end;
```

Theorem 1. *(i) A random polynomial of degree $n \geq 2$ in $\mathbb{F}_q[x]$ has a probability $1 - 1/q$ to be squarefree.*
(ii) The degree of the non-squarefree part of a random polynomial has expected value asymptotic to

$$C_q = \sum_{n \geq 1} \frac{n I_n}{q^{2n} - q^n},$$

and a geometrically decaying probability tail. We have $C_q \sim 1/q$ as $q \to \infty$.

PROOF. Part (i) is classical and is the consequence of Eq. (5). As for (ii), the bivariate generating function of the degree of the non-squarefree part of monic polynomials in $\mathbb{F}_q[x]$ is, by the symbolic methods of Section 3,

$$P(z, u) = \prod_{n \geq 1} \left(1 + \frac{z^n}{1 - u^n z^n}\right)^{I_n}.$$

The mean degree of the non-squarefree part is obtained from the derivative $P_u(z, 1)$ by singularity analysis. The generating function $P(z, 3/2)$ is dominated by $P(z)$ near its dominant singularity, so that the geometrically decaying probability tail holds. Finally, the asymptotic value of C_q as $q \to \infty$ is obtained by means of the expansion $n\, I_n = q^n + \mathcal{O}(q^{n/2})$. □

Theorem 1 has important consequences for the recursive structure of the `factor` procedure. First, the overall cost of the recursive calls (Step 4 in the top-level procedure) remains $\mathcal{O}(1)$ on average. Next, alternative strategies giving the full squarefree factorization [12, p.345] have asymptotically equivalent costs. Finally, the ERF phase has a cost dominated by its first gcd.

Theorem 2. *The expected cost of the ERF phase applied to a random polynomial of degree n is asymptotically that of a single gcd,*

$$\overline{\tau ERF}_n \sim \tau_2 n^2.$$

5. Distinct-degree factorization (DDF)

The second stage of our reference algorithm requires finding the *distinct-degree factorization* (DDF) of the squarefree polynomial a. This means expressing a in the form $b_1 \cdot b_2 \cdots b_n$ where b_k is the product of irreducible factors of degree k. The principle is that the polynomial $x^{q^k} - x \in \mathbb{F}_q[x]$ is the product of all monic irreducible polynomials in $\mathbb{F}_q[x]$ whose degree divides k (see [19], p. 91).

```
procedure DDF(a : polynomial);  [a is assumed squarefree]
        n := deg(a); g := a; h := x;
        for k := 1 to n do
1.            h := h^q mod g;
2.            b[k] := gcd(h-x,g);
3.            g := g/b[k]; [a without irred factors of deg<=k]
4.            if b[k] <> 1 then h := h mod g fi;
        od;
        return(b[1].b[2]...b[n]);
end;
```

The computation in step 1 is done by means of the classical *binary powering* method [16, p. 441-442]. With $\nu(q)$ the number of ones in the binary representation of q, the number of products needed to compute $h^q \pmod g$ is

$$\lambda(q) = \lfloor \log_2 q \rfloor + \nu(q) - 1. \tag{8}$$

By the exponential tail result of Thm 1, we need only consider the cost of DDF applied to the squarefree part a of the input polynomial f and our subsequent analyses are all relative to the statistics induced by a random input f of degree n.

Theorem 3. *The expected cost of the basic DDF phase satisfies*

$$\overline{\tau DDF}_n \sim \frac{5}{12}\left(\lambda(q)\tau_1 + \tau_2\right) n^3 \qquad \textit{where} \quad \lambda(q) = \lfloor \log_2 q \rfloor + \nu(q) - 1.$$

PROOF. The cost of the basic DDF is $C_1 + C_2 + C_3 + C_4$, where C_j denotes the cost of line number j. We let $\overline{C_j}$ be the expectation of C_j. Since the mean number of factors of f is $\mathcal{O}(\log n)$, we find $\overline{C_3} + \overline{C_4} = \mathcal{O}(n^2 \log n)$.

Let d_k denote the degree of polynomial g when the kth iteration of the main loop starts; the parameter d_k is also the sum of the degrees of the distinct factors

of f with degree $\geq k$. The quantity C_1+C_2 is equal to $(\lambda(q)\tau_1+\tau_2)\sum_{k\geq 1} d_k^2$. The bivariate generating function associated with d_k is, by the basic decompositions,

$$P_k(z,u) = \prod_{j<k}\left(\frac{1}{1-z^j}\right)^{I_j} \prod_{j\geq k}\left(1+u^j\frac{z^j}{1-z^j}\right)^{I_j}.$$

The expected value of $C = \sum_{k\geq 1} d_k^2$ is then given by

$$\overline{C} = \frac{1}{q^n}[z^n]R(z), \qquad R(z) = \sum_{k\geq 1}\left(\frac{\partial^2 P_k}{\partial u^2}(z,u) + \frac{\partial P_k}{\partial u}(z,u)\right)_{u=1}.$$

The GF $R(z)$ involves the coefficients I_n and from the main estimate $nI_n = q^n + \mathcal{O}(q^{n/2})$, the behaviour near the dominant singularity $z = 1/q$ results: $R(z) \sim \frac{5}{2}(1-qz)^{-4}$. Singularity analysis entails that $[z^n]R(z/q) \sim \frac{5}{12}n^3$. □

5.1. The "half-degree" rule. A natural idea is to stop the DDF loop when $k = n/2$, since at this stage the remaining factor is either 1 or it is irreducible.

Theorem 4. *The expected cost of the "half-degree rule" DDF phase satisfies*

$$\overline{\tau DDF}_n^{(HD)} \sim \frac{5}{16}\left(\lambda(q)\tau_1+\tau_2\right)n^3 \qquad \text{where} \quad \lambda(q) = \lfloor \log_2 q \rfloor + \nu(q) - 1.$$

PROOF. The cost is now given by $(\lambda(q)\tau_1+\tau_2)C^{(1)}$, where $C^{(1)} = \sum_{k\leq n/2} d_k^2$. Let D_1 be the highest degree of all irreducible factors of f. We study the difference $C^{(2)} = C - C^{(1)}$. If $D_1 \leq n/2$, we have $C^{(2)} = 0$, otherwise we have $C^{(2)} = (D_1 - \lfloor n/2 \rfloor)\, D_1^2$ since there can be only one factor of degree larger than $n/2$, namely D_1. Thus the mean value of $C^{(2)}$ is given by

$$\overline{C^{(2)}} = \sum_{n/2<k\leq n} \Pr(D_1 = k)\left(k - \left\lfloor \frac{n}{2}\right\rfloor\right) k^2. \tag{9}$$

The probability $\Pr(D_1 = k)$ is derived from the generating function $\chi_k(z)$ of polynomials whose factors have all degree $\leq k$ as

$$\Pr(D_1 = k) = \frac{1}{q^n}[z^n]\left(\chi_k(z) - \chi_{k-1}(z)\right), \qquad \chi_k(z) = \prod_{j=1}^{k}\left(\frac{1}{1-z^j}\right)^{I_j}.$$

When $k > n/2$, the n-th coefficient of $\chi_k(z) - \chi_{k-1}(z)$ is obtained from

$$\chi_k(z) - \chi_{k-1}(z) = P(z)\left(1-(1-z^k)^{I_k}\right)\prod_{j>k}(1-z^j)^{I_j} = P(z)\left(I_k z^k + \mathcal{O}(z^{n+1})\right)$$

which entails $\Pr(D_1 = k) = I_k/q^k \sim 1/k$ for $n/2 < k \leq n$. Plugging this information into (9) gives $\overline{C^{(2)}} \sim \frac{5}{48}n^3$, thus $\overline{C^{(1)}} = \overline{C} - \overline{C^{(2)}} \sim \frac{15}{48}n^3$. □

Thus, the half-degree rule results in a savings of 25% asymptotically.

5.2. The "early-abort" strategy. A still better strategy called "early abort" consists in stopping the main loop of DDF as soon as $2k$ exceeds the degree of the remaining factor, since then the remaining factor must be irreducible. The analysis now has some analogy to that of integer factoring given by Knuth and Trabb-Pardo [17].

Theorem 5. *The expected cost of the "early-abort rule" DDF phase satisfies*

$$\overline{\tau DDF}_n^{(EA)} \sim \delta\,(\lambda(q)\tau_1+\tau_2)\,n^3, \qquad \text{where} \quad \delta = 0.2668903307\ldots,$$

$$\delta = \frac{5}{12} - \frac{1}{3}\int_0^\infty e^{-2x}\exp\left(-\int_x^\infty \frac{e^{-u}}{u}\,du\right)\frac{1-x^2}{x}\,dx.$$

The constant δ is a close relative of the famous Golomb constant that intervenes in the expectation of the longest cycle in a random permutation [23].

PROOF. Let D_1 and D_2 be the degrees of the two irreducible factors of f of highest degree, setting $D_2 = 0$ if a is irreducible. The iteration is now aborted at step $k = \max\{\lfloor D_1/2\rfloor, D_2\} + 1$. The cost of DDF with this stopping rule becomes $C^{(3)} = \sum_{k\le \max\{\lfloor D_1/2\rfloor, D_2\}} d_k^2$ times the constant $(\lambda(q)\tau_1+\tau_2)$. Consider the difference $C^{(4)} = C - C^{(3)}$. We have

$$C^{(4)} = \begin{cases} (D_1 - \lfloor D_1/2\rfloor)\,D_1^2 & \text{if } D_1/2 > D_2 \\ (D_1 - D_2)D_1^2 & \text{if } D_1/2 \le D_2. \end{cases}$$

The generating function of polynomials for which $D_1 > 2D_2$ is given by

$$\phi_{D_1}(z) = \left[\prod_{1\le \ell < D_1/2}\left(\frac{1}{1-z^\ell}\right)^{I_\ell}\right] I_{D_1}\left(\frac{z^{D_1}}{1-z^{D_1}}\right),$$

and the generating function of polynomials for which $D_2 < D_1 \le 2D_2$ is given by

$$\psi_{D_1,D_2}(z) = \left[\prod_{1\le \ell < D_2}\left(\frac{1}{1-z^\ell}\right)^{I_\ell}\right]\left[\left(\frac{1}{1-z^{D_2}}\right)^{I_{D_2}} - 1\right]\left[I_{D_1}\frac{z^{D_1}}{1-z^{D_1}}\right]$$

(we do not need to take the case $D_1 = D_2$ into account since it contributes 0 to $C^{(4)}$). Hence, the GF of the cumulated values of the parameter $C^{(4)}$,

$$\Phi(z) = \sum_{D_1}\left(D_1 - \left\lfloor\frac{D_1}{2}\right\rfloor\right) D_1^2\,\phi_{D_1}(z) + \sum_{D_2 < D_1 \le 2D_2}(D_1 - D_2)D_1^2\,\psi_{D_1,D_2}(z).$$

The analysis of this generating function near its positive dominant singularity q^{-1} is done by approximating sums with integrals (Euler-Maclaurin summation) after the change of variables $z = e^{-t}q^{-1}$. A somewhat delicate analysis shows that $\Phi(z/q) \sim c_0(1-z)^{-4}$ as $z \to 1^-$, where $c_0 = \frac{5}{2} - 6\delta$. A Tauberian argument is needed since the positive singularity is not isolated. □

The global savings of the early abort rule is of 36% and the expected cost of $\mathcal{O}(\log q \cdot n^3)$ for DDF clearly dominates in the whole factorization chain.

6. The output configuration of DDF

The DDF procedure does not completely factor a polynomial that has different irreducible factors of the same degree. However, as shown by the following theorem, "most" of the factoring has been completed after DDF.

Theorem 6. *(i) The asymptotic probability of a complete DDF factorization is*

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

$c_2 \doteq 0.6656$, $c_{257} \doteq 0.5618$, $c_\infty = e^{-\gamma} \doteq 0.5614$, *where* γ *is Euler's constant.*
(ii) The expected degree of the part of the input polynomial subjected to the EDF phase is asymptotic to $\log n$.

PROOF. (i) The GF of polynomials with irreducible factors of distinct degrees

$$\prod_{n\geq 1}\left(1+I_n\frac{z^n}{1-z^n}\right) \tag{10}$$

has the equivalent form $(1-qz)^{-1}\,\phi(z)$, where $\phi(z)$ is obtained by multiplying each term of the product (10) by $(1-z^n)^{I_n}$. The function $\phi(z)$ is continuous at $1/q$ and a Tauberian-like argument applies. Finally, when q is large, the relation $nI_n = q^n + \mathcal{O}(q^{n/2})$ is used to prove that c_q tends to $\prod_{n\geq 1}(1+1/n)e^{-1/n} = e^{-\gamma}$.

(ii) The bivariate generating function associated to the total degree of the nontrivial part of DDF is

$$P(z,u) = \prod_{n\geq 1}\left[\left(1+u^n\frac{z^n}{1-z^n}\right)^{I_n} - (u^n-1)I_n\,\frac{z^n}{1-z^n}\right].$$

The corresponding mean value is $q^{-n}[z^n]R(z)$, where $R(z)$ equals $P_u(z,u)|_{u=1}$. Near $z=q^{-1}$, $R(z)$ behaves like $(1-qz)^{-1}\log(1-qz)^{-1}$. As before, a Tauberian-like argument is needed, giving $q^{-n}[z^n]R(z) \sim \log n$. □

7. Equal-degree factorization (EDF)

From Section 5, the factorization problem is eventually reduced to factoring a collection of polynomials b_j of a special form that have all their irreducible factors of the same (known) degree j. Our reference chain uses the classical Cantor-Zassenhaus algorithm [3]. The analysis combines a recursive partioning problem akin to digital tries [15] with estimates on the degree of irreducible factors of random polynomials [14].

```
procedure EDF(b : polynomial, k : integer);
[b is a product of irreducibles of degree k]
        if degree(b) <= k then return(b) fi;
        h := randpoly(degree(b)-1);
1.      a := h^((q^k-1)/2)-1 mod b;
2.      d := gcd(a,b);
        return(EDF(d,k).EDF(b/d,k));
end;
```

7.1. EDF and digital tries. By elementary properties of finite fields, each factor of b has a probability $\alpha = \frac{q-1}{2q}$ to be a factor of d and the complementary probability $\beta = \frac{q+1}{2q}$ to divide b/d. The probability that a random choice leads to a split of b that is of type $\langle \ell, j-\ell \rangle$ is thus the Bernoulli probability $\binom{j}{\ell}\alpha^\ell \beta^{j-\ell}$.

Theorem 7. *The cost of the EDF algorithm on polynomials with j irreducible factors of degree k is $C_{j,k} =$*

$$\left(\frac{1}{2\alpha\beta} j(j-1) + j \sum_{m\geq 0} \sum_{\ell=0}^{m} \binom{m}{\ell} \alpha^{m-\ell}\beta^\ell \left(1 - (1-\alpha^{m-\ell}\beta^\ell)^{j-1}\right) \right) \ (\mu_k \tau_1 + \tau_2) k^2,$$

where $\mu_k = \lambda((q^k-1)/2) = \left\lfloor \log_2 \frac{q^k-1}{2} \right\rfloor + \nu\left(\frac{q^k-1}{2}\right) - 1$.

PROOF. A complete recursive execution of the EDF procedure is equivalent to developing a binary tree of possibilities. For a tree t with root subtrees t_0, t_1, we thus consider a general cost function of the additive type,

$$C[t] = e_{|t|} + C[t_0] + C[t_1]. \tag{11}$$

where $e_{|t|}$ is a (problem specific) "toll" function that depends on the size $|t|$ (number of nonempty external nodes) of t.

Like for tries [15], the subtree sizes obey the Bernoulli probability given above. Thus, the expectation c_j of $C[t]$ over trees of size j satisfies the recurrence

$$c_j = e_j + \sum_{\ell=0}^{j} \binom{j}{\ell} \alpha^\ell \beta^{j-\ell} (c_\ell + c_{j-\ell}) = e_j + \sum_{\ell=0}^{j} \binom{j}{\ell} (\alpha^\ell \beta^{j-\ell} + \alpha^{j-\ell}\beta^\ell) c_\ell.$$

This translates, in terms of exponential generating functions, $C(z) = \sum_j c_j z^j/j!$ and $E(z) = \sum_j e_j z^j/j!$, into the functional equation $C(z) = E(z) + e^{\beta z} C(\alpha z) + e^{\alpha z} C(\beta z)$, that iterates to give the explicit solution

$$C(z) = \sum_{j\geq 0} \sum_{\ell=0}^{j} \binom{j}{\ell} E(\alpha^{j-\ell}\beta^\ell z) e^{z(1-\alpha^{j-\ell}\beta^\ell)}. \tag{12}$$

Here, the toll function $e_j = j^2 - \delta_{j,1}$ leads to

$$C(z) = \frac{1}{2\alpha\beta} z^2 e^z + z \sum_{m\geq 0} \sum_{\ell=0}^{m} \binom{m}{\ell} \alpha^{m-\ell}\beta^\ell \left(e^z - e^{z(1-\alpha^{m-\ell}\beta^\ell)}\right),$$

by means of Eq. (12). From there, an explicit expression for the coefficients results. The analysis is completed by finally taking into account the cost of multiplications modulo b that intervene in the computation of $h^{(q^k-1)/2} \mod b$ by the binary powering algorithm, leading to the μ_k. □

7.2. Complete analysis. Completing the analysis of EDF only requires weighting the costs given by Theorem 7 by the probability $\Pr(\omega_n(k) = j)$ of finding j irreducible factors of degree k. Let $\omega_n(k)$ be the random variable counting the number of distinct irreducible factors of degree k in a random polynomial of degree n. The corresponding probability distribution can be computed by the decomposition techniques of Section 3, see [14], and one has:

$$\Pr\{\omega_n(k) = j\} = \begin{cases} \dfrac{\binom{I_k}{j}}{q^{kj}} (1 - q^{-k})^{I_k - j} \underset{q\to\infty}{\sim} e^{-1/k} \dfrac{k^{-j}}{j!} & \text{if} \quad n \geq kI_k, \\ \dfrac{\binom{I_k}{j}}{q^{kj}} \displaystyle\sum_{\ell=0}^{\lfloor n/k \rfloor - j} (-1)^\ell \frac{\binom{I_k - j}{\ell}}{q^{k\ell}} & \text{if} \quad kj \leq n < kI_k. \end{cases}$$

The distribution is essentially a negative binomial that can be approximated by a Poisson law of parameter $1/k$. Hence:

Theorem 8. *The expected cost of the EDF phase satisfies*

$$\overline{\tau EDF}_n \sim \frac{\tau_1}{\alpha\beta} \sum_{k=1}^{\lceil n/2 \rceil} \mu_k, \qquad \mu_k = \left\lfloor \log_2 \frac{q^k - 1}{2} \right\rfloor + \nu\left(\frac{q^k - 1}{2}\right) - 1.$$

In addition, this cost is $\mathcal{O}(n^2)$ and

$$(13)\ \overline{\tau EDF}_n \sim \left(\frac{3}{4}\tau_1 \frac{q^2}{q^2 - 1} \log_2 q \cdot n^2\right)(1 + \xi_n + o(1)), \qquad -\frac{1}{3} \leq \xi_n \leq \frac{1}{3}.$$

PROOF. The intuition behind the proof is that the major contribution comes from situations where just 2 factors are present, the other cases having globally a very small probability of occurrence. Let $\overline{\mathrm{E}_k}$ be the expected value of the cost of the EDF algorithm corresponding to degree k. By definition, we have $\overline{\mathrm{E}_k} = \sum_{j\geq 2} \Pr(\omega_n(k) = j)\, C_{j,k}$, where $C_{j,k}$ is given by Theorem 7.

First, the form of the distribution of the number of distinct factors implies

$$\Pr(\omega_n(k) = 2) = \frac{\binom{I_k}{2}}{q^{2k}}(1 + \mathcal{O}(1/k)) \quad \text{for } 2k \leq n, \quad \Pr(\omega_n(k) = j) = O\left(\frac{1}{j! k^j}\right).$$

Next, from Theorem 7, we deduce

$$C_{0,k} = C_{1,k} = 0, \quad C_{2,k} = \frac{2}{\alpha\beta}(\mu_k \tau_1 + \tau_2)\, k^2, \quad \text{and uniformly} \quad C_{j,k} = \mathcal{O}(j^2 k^3).$$

This entails that, as $k \to \infty$ with $2k \leq n$,

$$\overline{\mathrm{E}_k} = C_{2,k} \frac{\binom{I_k}{2}}{q^{2k}}(1 + \mathcal{O}(1/k)) + \sum_{j\geq 3} \mathcal{O}\left(\frac{k^{-j}}{j!} j^2 k^3\right) = \frac{\tau_1}{\alpha\beta}\mu_k + \mathcal{O}(1),$$

while $\overline{\mathrm{E}_k} = 0$ for $2k > n$. Thus, the overall cost of the EDF component is $\sum_k \overline{\mathrm{E}_k} = \frac{\tau_1}{\alpha\beta} \sum_{k=1}^{\lceil n/2 \rceil} \mu_k + \mathcal{O}(n)$. The second form is easily obtained from $k \log_2 q - 2 \le \mu_k \le 2k \log_2 q$. □

Under the unproven assumption that the binary representation of q^k behaves like that of a random integer, the arithmetic function ξ_n should be close to 0.

Acknowledgement. Work of P.F. and X. G. has been supported by the Long Term Research Project Alcom-IT (# 20244) of the European Union.

References

1. BERLEKAMP, E. R. *Algebraic Coding Theory*. Mc Graw-Hill, 1968. Revised edition, 1984.
2. BUCHMANN, J. Complexity of algorithms in number theory. In *Number Theory: Proceedings of the First Conference of the Canadian Number Theory Association* (1990), Walter de Gruyter, pp. 37–53.
3. CANTOR, D. G., AND ZASSENHAUSS, H. A new algorithm for factoring polynomials over finite fields. *Mathematics of Computation 36* (1981), 587–592.
4. CAR, M. Factorisation dans $F_q[x]$. *Comptes-Rendus de l'Académie des Sciences 294 (Ser. I)* (1982), 147–150.
5. CARLITZ, L. The arithmetic of polynomials in a Galois field. *American Journal of Mathematics 54* (1932), 39–50.
6. CHOR, B., AND RIVEST, R. A knapsack type public key cryptosystem based on on arithmetics over finite fields. *IEEE Transactions on Information Theory 34* (1988), 901–909.
7. COMTET, L. *Advanced Combinatorics*. Reidel, Dordrecht, 1974.
8. DEDEKIND, R. Abriss einer Theorie der höhern Congruenzen in Bezug auf einen reellen Primzahlmodulus. *Journal für die reine und angewandte Mathematik 54* (1857), 1–26.
9. FLAJOLET, P., AND ODLYZKO, A. M. Singularity analysis of generating functions. *SIAM Journal on Discrete Mathematics 3*, 2 (1990), 216–240.
10. FLAJOLET, P., AND SORIA, M. Gaussian limiting distributions for the number of components in combinatorial structures. *Journal of Combinatorial Theory, Series A 53* (1990), 165–182.
11. GAUSS, C. F. *Untersuchungen über höhere Mathematik*. Chelsea, New York, 1889.
12. GEDDES, K. O., CZAPOR, S. R., AND LABAHN, G. *Algorithms for Computer Algebra*. Kluwer Academic Publishers, Boston, 1992.
13. GREENE, D. H., AND KNUTH, D. E. *Mathematics for the analysis of algorithms*, second ed. Birkhauser, Boston, 1982.
14. KNOPFMACHER, J., AND KNOPFMACHER, A. Counting irreducible factors of polynomials over a finite field. *Discrete Mathematics 112* (1993), 103–118.
15. KNUTH, D. E. *The Art of Computer Programming*, vol. 3: Sorting and Searching. Addison-Wesley, 1973.
16. KNUTH, D. E. *The Art of Computer Programming*, 2nd ed., vol. 2: Seminumerical Algorithms. Addison-Wesley, 1981.
17. KNUTH, D. E., AND PARDO, L. T. Analysis of a simple factorization algorithm. *Theoretical Computer Science 3* (1976), 321–348.
18. LENSTRA, H. W. On the Chor Rivest cryptosystem. *Journal of Cryptology 3* (1991), 149–155.
19. LIDL, R., AND NIEDERREITER, H. *Finite Fields*, vol. 20 of *Encyclopedia of Mathematics and its Applications*. Addison-Wesley, 1983.
20. ODLYZKO, A. M. Discrete logarithms and their cryptographic significance. In *Advances in Cryptology* (1985), Lecture Notes in Computer Science, Springer Verlag, pp. 224–314.
21. ODLYZKO, A. M. Asymptotic enumeration methods. In *Handbook of Combinatorics*, M. G. R. Graham and L. Lovász, Eds., vol. II. Elsevier, Amsterdam, 1995, pp. 1063–1229.
22. POSTNIKOV, A. G. *Tauberian theory and its applications*, vol. 144 of *Proceedings of the Steklov Institute of Mathematics*. American Mathematical Society, 1980.
23. SHEPP, L. A., AND LLOYD, S. P. Ordered cycle lengths in a random permutation. *Transactions of the American Mathematical Society 121* (1966), 340–357.
24. SHOUP, V. A new polynomial factorization algorithm and its implementation. Preprint, 1994.