Some observations to speed the polynomial selection in the number field sieve

Min Yang [a], Qingshu Meng [b], Zhangyi Wang [b], Huanguo Zhang [b]

[a] *International School of Software, Wuhan University, Hubei, China, 430072*

[b] *Computer School, Wuhan University, Hubei, China, 430072*

Abstract

If the yield of a polynomial pair is closely correlated with the coefficients of the polynomial pair, we can select polynomials by checking the coefficients first. This can speed the selection of good polynomials. In this paper, we aim to study the correlation between the polynomial coefficients and the yield of the polynomials. By heuristic analysis and some experiments, we find that the yield of polynomial with the ending coefficient containing many small primes is usually better than the one whose ending coefficient does not contain. The ending coefficient has closer correlation with the yield than the leading coefficient has. The number of real roots can be determined only by partial coefficients of the polynomial if it is skewed. All these observations can be used to speed the search of good polynomials for the number filed sieve.

Key words: integer factorization, number field sieve, polynomial selection, coefficients

1 Introduction

The general number field sieve is known as the asymtotically fastest algorithm for factoring large integers. It is based on the observation that if $a^2 = b^2 \, mod \, N$ and $a \neq b$, $gcd(a-b, N)$ will give a proper factor of N with at least a half chance. The number field sieve starts by choosing two irreducible and coprime polynomials f(x) and g(x) over Z which share a common root m modulo N. Let $F(x,y) = y^{d_1} f(x/y)$ and $G(x,y) = y^{d_2} g(x/y)$ be the homogenized polynomials

* Corresponding author
Email address: yangm75@hotmail.com (Min Yang).

corresponding to $f(x)$ and $g(x)$, where d_1 and d_2 are the degree of $f(x)$ and $g(x)$ respectively. We want to find many coprime pairs $(a, b) \in Z^2$ such that the polynomials values $F(a, b)$ and $G(a, b)$ are simultaneously smooth with respect to some upper bound B and the pair (a, b) is called a relation. An integer is smooth with respect to bound B (or B-smooth) if none of its prime factors are larger than B. If we find enough number of relations, we can construct:

$$\prod_{(a,b)\in S} (a - b\alpha_1) = \beta_1^2, where\ f(\alpha_1) = 0, \beta_1 \in Z[\alpha_1]$$

$$\prod_{(a,b)\in S} (a - b\alpha_2) = \beta_2^2, where\ g(\alpha_2) = 0, \beta_2 \in Z[\alpha_2].$$

As there exists $\varphi_1(\alpha_1) = m\ mod\ N$ and $\varphi_2(\alpha_2) = m\ mod\ N$, we have $\varphi_1(\beta_1^2) = \varphi_2(\beta_2^2)$ If we let $\varphi_1(\beta_1) = x$ and $\varphi_2(\beta_2) = y$, then $y^2 = x^2\ mod\ N$, and we have constructed a congruent squares and so may attempt to factor N by computing $gcd(x - y, N)$.

In order to obtain enough relations, selecting a polynomial with high probability of being smooth is very important. A good polynomial not only can decrease sieving time, but also can reduce the expected matrix size[7]. The polynomial selection is now a hot research area. Based on base-m method and with translate and rotate technique[7], non-skewed or skewed polynomial pair can be constructed, where one polynomial $f(x)$ is nonlinear and the other $g(x)$ is monic and linear. If the linear polynomial is nonmonic, the size of nonlinear polynomial can be greatly reduced[3,1]. The two methods above are called linear method. Montgomery[6] proposed the nonlinear method, where the two polynomials are both nonlinear. Recently several papers[4,8,9] address nonlinear polynomials construction problem. In this paper, we don't mean to propose new polynomial construction method, but to study the correlation between the polynomial coefficients and the yield of the polynomials. If they are closely related, we can select polynomial by checking the coefficients first. This takes less time and would speed the selection of polynomials.

The paper is organized as follows. In Section 2 we review elements related to the yield of a polynomial. In Section 3 we recite the number of real roots of a rational polynomial. In Section 4 we analysis the effect of the ending coefficient and leading coefficient on the yield respectively. In Section 5 we analysis the effects of coefficients on the number of real roots and on the yield. Finally we make a conclusion in Section 6.

2 Elements related to smoothness of a polynomial

An integer is said to be B-smooth if the integer can be factored into factors bounded by B. By Dickman function, given the smooth bound B, the less the integer is, the more likely the integer is B-smooth. In number field sieve, we want the homogenous form $F(x,y) = a_d x^d + \cdots + a_1 x y^{d-1} + a_0 y^d$ of the polynomial $f(x) = a_d x^d + \cdots + a_1 x + a_0$ to be small. In [7], the size and root property are used to describe the quantity. By size we refer to the magnitude of the values taken by $F(x,y)$. By root property we refer to the distribution of the roots of $F(x,y)$ modulo small p^k, for p prime and $k \geq 1$. If $F(x,y)$ has many roots modulo small p^k, values taken by $F(x,y)$ "behave" as if they are smaller than they actually are. That is, on average, the likelihood of $F(x,y)$ values being smooth is increased. It has always been well understood that size affects the yield of $F(x,y)$. In [2], the number of real roots, the order of Galois group of $f_1(x)f_2(x)$ were taken into account. By the number of real roots, if a/b is near a real root, the value $F(a,b)$ will be small and will be smooth with high chance. By the order of Galois group of $f_1 f_2$, it is better to chose polynomial for which the order of Galois group of $f_1 f_2$ are small, because they provide more free relations.

Obviously, if the coefficients of $f(x)$ are small, $F(x,y)$ would have good size property. In order to obtain polynomial with small coefficients, we can search extensively, or let the linear polynomial be nonmonic as suggested in [**?**,3]. In order to obtain good root property, usually it is required that the leading coefficient contains many small prime as its factors[7]. As for number of the real roots, it is left as random.

3 The number of real roots of a polynomial

In [10,11],the number of real roots or roots distribution of a rational polynomial is given by CDS(complete discrimination system).

In degree 3, take polynomial $f(x) = ax^3 + bx^2 + cx + d$ as example. The CDS is

$$\Delta = 18abcd - 4b^3 d + b^2 c^2 - 4ac^3 - 27a^2 d^2.$$

The root distribution is as follows.

(1) If $\Delta > 0$, the equation has three distinct real roots.
(2) If $\Delta = 0$, the equation has a multiple root and all its roots are real.
(3) If $\Delta < 0$, the equation has one real root and two nonreal complex conjugate roots.

In degree 4, take $f(x) = a_0x^4 + a_1x^3 + a_2x^2 + a_3x + a_4, (a_0 \neq 0)$ as example. Its CDS is as follows:

$$D_2 = 3a_1^2 - 8a_2a_0,$$

$$D_3 = 16a_0^2a_4a_2 - 18a_0^2a_3^2 - 4a_0a_2^3 + 14a_0a_3a_1a_2 - 6a_0a_4a_1^2 + a_2^2a_1^2 - 3a_3a_1^3,$$

$$D_4 = 256a_0^3a_4^3 - 27a_0^2a_3^4 - 192a_0^2a_3a_4^2a_1 - 27a_1^4a_4^2 - 6a_0a_1^2a_4a_3^2 + a_2^2a_3^2a_1^2 - 4a_0a_2^3a_3^2 + 18a_2a_4a_1^3a_3 + 144a_0a_2a_4^2a_1^2 - 80a_0a_2^2a_4a_1a_3 + 18a_0a_2a_3^3a_1 - 4a_2^3a_4a_1^2 - 4a_1^3a_3^3 + 16a_0a_2^4a_4 - 128a_0^2a_2^2a_4^2 + 144a_0^2a_2a_4a_3^2,$$

$$E = 8a_0^2a_3 + a_1^3 - 4a_0a_1a_2.$$

The following table gives the numbers of real and imaginary roots and multiplicities of repeated roots in all cases:

(1)	$D_4 > 0 \bigwedge D_3 > 0 \bigwedge D_2 > 0$	$\{1,1,1,1\}$
(2)	$D_4 > 0 \bigwedge (D_3 \leq 0 \bigvee D_2 \leq 0)$	$\{\}$
(3)	$D_4 < 0$	$\{1,1\}$.
(4)	$D_4 = 0 \bigwedge D_3 > 0$	$\{2,1,1\}$
(5)	$D_4 = 0 \bigwedge D_3 < 0$	$\{2\}$
(6)	$D_4 = 0 \bigwedge D_3 = 0 \bigwedge D_2 > 0 \bigwedge E = 0$	$\{2,2\}$
(7)	$D_4 = 0 \bigwedge D_3 = 0 \bigwedge D_2 > 0 \bigwedge E \neq 0$	$\{3,1\}$
(8)	$D_4 = 0 \bigwedge D_3 = 0 \bigwedge D_2 < 0$	$\{\}$
(9)	$D_4 = 0 \bigwedge D_3 = 0 \bigwedge D_2 = 0$	$\{4\}$

where the right column of the table describes the situations of the roots. For example, $(1, 1, 1, 1)$ means four real simple roots and $(2, 1, 1)$ means one real double root plus two real simple roots.

In degree 5, take $f(x) = x^5 + px^3 + qx^2 + rx + s$ as example. Its CDS is as follows:

$$D_2 = -p$$

$$D_3 = 40rp - 12p^3 - 45q^2$$

$$D_4 = 12p^4r - 4p^3q^2 + 117prq^2 - 88r^2p^2 - 40qp^2s + 125ps^2 - 27q^4 - 300qrs + 160r^3$$

$$D_5 = -1600qsr^3 - 3750ps^3q + 2000ps^2r^2 - 4p^3q^2r^2 + 16p^3q^3s - 900rs^2p^3 + 825q^2p^2s^2 + 144pq^2r^3 + 2250q^2rs^2 + 16p^4r^3 + 108p^5s^2 - 128r^4p^2 - 27q^4r^2 + 108q^5s + 256r^5 + 3125s^4 - 72p^4rsq + 560r^2p^2sq - 630prq^3s$$

$$E_2 = 160r^2p^3 + 900q^2r^2 - 48rp^5 + 60q^2p^2r + 1500rpsq + 16q^2p^4 - 1100qp^3s + 625s^2p^2 - 3375q^3s$$

$$F_2 = 3q^2 - 8rp$$

The following table gives the numbers of real and imaginary roots and multiplicities of repeated roots of polynomial in all cases:

(1)	$D_5 > 0 \bigwedge D_4 > 0 \bigwedge D_3 > 0 \bigwedge D2 > 0$	$\{1,1,1,1,1\}$
(2)	$D_5 > 0 \bigwedge (D_4 \leq 0 \bigvee D_3 \leq 0 \bigvee D_2 \leq 0)$	$\{1\}$
(3)	$D_5 < 0$	$\{1,1,1\}$

(4) $D_5 = 0 \bigwedge D_4 > 0$ $\{2,1,1,1\}$
(5) $D_5 = 0 \bigwedge D_4 < 0$ $\{2,1\}$
(6) $D_5 = 0 \bigwedge D_4 = 0 \bigwedge D_3 > 0 \bigwedge E \neq 0$ $\{2,2,1\}$
(7) $D_5 = 0 \bigwedge D_4 = 0 \bigwedge D_3 > 0 \bigwedge E = 0$ $\{3,1,1\}$
(8) $D_5 = 0 \bigwedge D_4 = 0 \bigwedge D_3 < 0 \bigwedge E \neq 0$ $\{1\}$
(9) $D_5 = 0 \bigwedge D_4 = 0 \bigwedge D_3 < 0 \bigwedge E = 0$ $\{3\}$
(10) $D_5 = 0 \bigwedge D_4 = 0 \bigwedge D_3 = 0 \bigwedge D_2 \neq 0 \bigwedge F_2 \neq 0$ $\{3,2\}$
(11) $D_5 = 0 \bigwedge D_4 = 0 \bigwedge D_3 = 0 \bigwedge D_2 \neq 0 \bigwedge F_2 = 0$ $\{4,1\}$
(12) $D_5 = 0 \bigwedge D_4 = 0 \bigwedge D_3 = 0 \bigwedge D_2 = 0$ $\{5\}$

4 The yield and the ending coefficient

Let $f(x) = a_d x^d + \cdots + a_1 x + a_0$ be a nonlinear polynomial. Let $F(x,y)$ be the homogenous form of polynomial f(x). Let p_q be the number of roots of the homogeneous polynomial F modulo p and let

$$\alpha(F) = \sum_{small\,prime\,p} (1 - p_q \frac{p}{p+1}) \frac{p}{p-1}.$$

In order to make $\alpha(F)$ small, we can increase the value of p_q. Equation $F(x,y) = 0 \; mod\, p$ has three kind of roots.

(1) If $p|a$ and $p|a_0$, the pair (a,b) is called the zero root.
(2) If $p|b$ and $p|a_d$, the pair (a,b) is called the projective root.
(3) The rest of pairs (a,b) satisfying $F(a,b) = 0\, mod\, p$ are called ordinary roots, or simply roots.

Correspondingly, there are three ways to increase the value of p_q. It is already known that the leading coefficient a_d containing many small primes can increase the number of projective roots. For example, the leading coefficient usually is the multiple of 60[3]. We propose if the ending coefficient contains many small primes, the number of zero roots can be increased. We will analyze soon. As for the ordinary roots, we don't know how to increase it.

Suppose the sieving area be $2A * B$. As $A/B \approx (a_0/a_d)^{\frac{1}{d}}$ and f(x) is usually skewed with $a_d << a_0$, A is much big than B. The number of case $p|a$ is about $2A/p$ and the number of $p|b$ is about B/p. Therefore, the number of pair (a,b) satisfying the first case is more than the one in second case.

In order to check whether the above heuristic analysis is right, we do many experiments. In our experiments, we let N be an integer about 30 digits. In experiment 1, the polynomials are generated by base-m method as described in [7], but without the optimization step.

Table 1
The trend of three parameter(256 rows/block)

Block num	1	2	3	4	5	6	7	8	9
num_{ad}	860	742	656	639	622	622	624	685	810
num_{a0}	286	354	410	466	484	532	633	659	723
num_{root}	448	430	450	440	458	478	486	482	524

Experiment 1:

(1) Generate polynomial as [7]. For each leading coefficient a_d below a bound, we examine

$$m \approx \lfloor (\frac{N}{a_d})^{\frac{1}{d}} \rfloor.$$

Check the magnitude of a_{d-1}, and of a_{d-2} compared to m, by computing the integral and non-integral parts of

$$\frac{N - a_d m^d}{m^{d-1}} = a_{d-1} + \frac{a_{d-2}}{m} + O(m^{-2}).$$

If these are sufficiently small, accept a_d and m, and we get a polynomial f(x) by the expansion of N with base m and leading coefficient a_d.

(2) Collect relations. For each above polynomial, skew the sieving area with skewness=$(\frac{a_0}{a_d})^{\frac{1}{d}}$. Randomly choose enough pair of coprime (a,b) in sieving area and check if they form a relation. Here we allow one large prime only for rational side, not the algebraic side[5]. For each polynomial,we denote the number of relations by num_{rel}.

(3) For each polynomial, there is a row corresponding to it. It includes the following items: the number of relations num_{rel}, the number of small primes contained in a_d below a predefined bound, denoted by num_{ad}, the number of small primes in a_0 below a predefined bound, denoted by num_{a0}. A file is formed.

(4) Sort the above file in ascending order with num_{rel} as key word. From the sorted file, we can find the parameter num_{a0} is also in ascending order, but not strictly. Parameter num_{ad} seems to be in big U shape, also not strictly. In order to obtain an obvious impression, we divide the sorted file by rows into many length-equal parts, each of which contains equal number of rows and calculate the sum of the parameters num_{ad} and num_{a0} in each part respectively.

Table 1 lists the sum of num_{ad} and num_{a0} respectively, where the parameters are as follows. $N =$ 39327284784436337729633(an integer in example 3 [4]). The degree of the nonlinear polynomial is 3. Sieving area is $2A \times A$, where $A =$ 4000, coprime pair (a, b) are chosen randomly from sieving area in a way like "for(a=-$A \times s$;a< $A \times s$;a+=rand()%6+1) for(b=1;b< A/s;b+=rand()%6+1), where $s = \sqrt[6]{a_0/a_3}$". From Table 1 we find that the ending coefficients correlate with the yields of polynomials more closely than the leading coefficient does. For polynomial of degree 4 or 5, we get similar results.

For nonmonic linear polynomial generated as suggested in [1,3], we can get similar results. For nonlinear polynomials as suggested in[4], we don't do the experiments. We conjecture the results should be similar.

By the analysis above and the experiments, we have:

Observation 1: Increasing the number of small primes which are contained in the ending coefficient as factors may increase the yield. The ending coefficient correlate more closely with the yield of the polynomial pair than the leading coefficient does.

In [7], it is said that computing the ideal decomposition for ideals corresponding to projective roots requires more effort than those corresponding to non-projective roots. Therefore, increasing zero roots is a better choice.

5 The number of real roots and the coefficients

A polynomial with more real roots are preferable in number field sieve because if a/b is near a real root, the value $F(a,b)$ will be small and will be smooth with high possibility. Usually the number of real roots is left as random in polynomial generation. In [10,11],the number of real roots or roots distribution of a rational polynomial is given by CDS(complete discrimination system). From CDS, the number of real roots should depends on all coefficients of the polynomial. However, the polynomial for NFS are skewed, not randomly generated. The number of real roots is correlated closely with partial polynomial coefficients. This can be used to polynomial selection.

In degree 3, from the expression of Δ, if the variable b or c is small enough, the $\Delta > 0$, which means the polynomial $f(x)$ has 3 real roots. If the polynomial is skewed, it is likely that the coefficient c of degree 1 will determine the number of real roots. The result of Experiment 2 coincide with this analysis. For degree 4, from the expression of D_4, the exponent of a_1, a_2, a_3 are 4. If the polynomial are skewed, the coefficient a_3 will determine that the number of real roots is 2. From the expression of D_3 and D_2, the coefficient a_2 play the key role. To obtain 4 real roots, the coefficient a_2 should be small enough and the absolute value of a_3 should be of similar size with that of a_2. The result of Experiment 2 coincide with the analysis too because in Experiment 2 the absolute values of a_2 and a_3 are of similar size. For degree =5, from the expression of CDS, it is complex to determine the number of real roots just from partial coefficients. However, in order to avoid obtaining only one real root, p should be small and negative such that all of D_4, D_3, D_2 are positive. In Experiment 2 we obtain only a few cases with 5 real roots.

The analysis above in case of degree 3 should be useful in choosing polynomial in nonlinear method, where a polynomial with degree 3 is already enough for practical purpose.

Experiment 2:

(1) Generate polynomial as step 1 of experiment 1.
(2) Collect relation as step 2 of experiment 1. Denote the number of relation by num_{rel}.
(3) For each polynomial, there is a row corresponding to it. The row has num_{rel}, the number of real roots num_{root}, and all coefficients as its items. We form a file now.
(4) Sort the above file in ascending order with num_{root} as the key word. From the sorted file, we observe the correlation between num_{root} and the polynomial coefficients. As Table 2 indicate, in case degree =3, we find the parameter num_{root} be determined almost only by the coefficient of degree 1. That is, if the coefficients of degree 1 is below some value, then the number of root is 3 for most cases. In case degree=4, the related coefficient is of degree=2, not degree =3. In case degree=5, no obvious phenomenon is observed.
(5) Sort the above file in ascending order with num_{rel} as the key word. From the sorted file, we observe the correlation between num_{root} and num_{rel}. We can find num_{root} is also in ascending order, but not strictly. In order to obtain an obvious impression, we divide the sorted file by rows into many length-equal parts, each of which contains equal number of rows and calculate the sum of the parameters num_{root} in each part.

Table 1 lists the sum of num_{root}, where the parameters are the same as in experiment 1. From table 1 we find that increasing the number of roots can increase the yield in degree 3. For degree 4 or 5, we get similar results, but not strong as the case in degree 3.

For polynomial generated as suggested by Kleinjung in [3], where the linear polynomial is nonmonic, the results is similar. As for the nonlinear polynomial, we don't do the experiments, but we conjecture results should be similar if the polynomials are skewed.

By the analysis above and the experiments, we have:

Observation 2: The number of real roots can be determined almost only by one coefficient in degree 3, be determined by two coefficients of the polynomial in degree 4. Increasing the number of real roots also increase the yield of the polynomial pair.

Table 2
number of root and the coefficients

polynomial degree	*degree of the related coefficient*	*number of real roots*
$degree = 3$	$degree = 1$	3
$degree = 4$	$degree = 2$	4
$degree = 5$	*unknown*	*unknown*

6 Conclusion

Studying the correlation between the yield of a polynomial and its coefficients is important because it take less computation if we can choose polynomial by checking its coefficients first. In this paper, we study the relation between the yield of a polynomial and its coefficients. The heuristic analysis and the experiments coincide well. They both show that the ending coefficient has more closely relationship with the yield of the polynomial than the leading coefficient has and the polynomial with the ending coefficient containing many small primes is more preferable. And the number of real roots can be determined almost only by one or two coefficients of the polynomial in degree 3 and 4. Increasing the number of real roots can increase the yield of the polynomial pair. All these observations can be used to speed the search of a good polynomial.

References

[1] J.P. BUHLER, H.W. LENSTRA, JR., C. POMERANCE, Factoring Integers With The Number Field Sieve, in A. K. Lenstra and H. W. Lenstra, Jr. (eds.),The Development of the Number Field Sieve, LNCS 1554, 1993, 50-94.

[2] M. Elkenbracht-Huizing,An Implementation of the Number Field Sieve, Experimental mathematics, Vol.5, No.3,1996,231-251.

[3] T. Kleinjung, On Polynomial Selection For The General Number Field Sieve, Mamathematics of Computtion, Vol.75, No.256, 2006, 2037C2047.

[4] N. Koo, G.H. Jo, and S. Kwon, On Nonlinear Polynomial Selection and Geometric Progression (mod N) for Number Field Sieve. https://eprint.iacr.org/2011/292.pdf

[5] A. K. LENSTRA, M. S. MANASSE, Factoring With Two Large Primes, Mathematics of computation, vol.63, No.208, 1994, 785-798.

[6] P. L. Montgomery, Small Geometric Progressions Modulo n, manuscript (1995).

[7] B. Murthy, Polynomial Selection for the Number Field Sieve Integer Factorisation Algorithm, Ph.D. thesis, The Australian National University, 1999.

[8] T. Prest, P. Zimmermann, Non-linear Polynomial Selection For The Number Field Sieve, Journal of Symbolic Computation, Vol.47, Issue 4, 2012, 401C409.

[9] R. S. Williams, Cubic Polynomials in the Number Field Sieve, Master Thesis, Texas Tech University, 2010.

[10] L. YANG, Recent Advances on Determining the Number of Real Roots of Parametric Polynomials, Journal of Symbolic Computation, (1999)Vol. 28, 225-242.

[11] http://en.wikipedia.org/wiki/Cubic_function,