Space Complexity Analysis of Sieving in the Number Field Sieve Integer Factorization

Qi Wang[1,a], Hongyan Zang[1,b], Xiubin Fan[2,c], Yu Wang[2,d]

[1]Department of Mathematics and Physics, Beijing University of Science and Technology, Beijing, 100083, China

[2]State Key Laboratory of Cryptology, Beijing, 100878, China

[a]email: wangqi1207bk@163.com, [b]email: a9801255@ustb.edu.cn,

[c]email: fanxiubin1966@sina.com , [d]email: sxuwy@hotmail.com

Keywords: number field sieve, integer factorization, mathematical expectation, $p-adic$ evaluation, space complexity

Abstract. The general number sieve is the most efficient algorithm known integer factorization, it consists of polynomial selection, sieving, solving equations and finding square roots. In this paper, the $p-adic$ evaluation provided by each root and the expected $p-$value are given, then we get the space complexity of sieving over the ring $\mathbb{Z}/2\mathbb{Z}$.

Introduction

In 1976, Diffie and Hellman published their paper "New direction in cryptography"[1]. It is regarded as a milestone for the research and development of cryptography. In their paper they first introduced public-key cryptography, also named asymmetric cryptography. Since then, public-key cryptography has been widely applied in encryption, digital signature, key exchange, and so on.

In 1977, Rivest, Shamir and Adleman proposed a public-key cryptographic algorithm suitable for both signing and encryption, known as RSA [2]. The RSA algorithm depends on assumed difficulty of the large integer factorization problem.

Pollard described a new method for factoring integers of a special form , the manuscript was enclosed with a letter to Odlyzko, dated 31 August 1988. This method is called the special number field sieve (SNFS)[3].

The general number field sieve (GNFS) was developed from the special number field sieve (SNFS) . It is the most efficient algorithm known integer factorization. GNFS has been used in many (current and previous) record factorization such as RSA-768[4]. In this paper, the term "number field sieve" refers to the general number field sieve unless otherwise mentioned. It consists of several stages as follows:

Step1: polynomial selection [5][6][7].

Let n be the integer to be factored. The number field sieve starts by choosing two irreducible and coprime polynomials $f(x)$ and $g(x)$ over $\mathbb{Z}$ which share a common root m modulo n. It is desirable that the polynomial pair can produce many smooth integer across the sieve region.

We assume (f,g) be the chosen polynomial pair. For convenience, $f(x)$ is referred to as the algebraic polynomial and $g(x)$ is referred to as the rational polynomial.

Setp2: sieving [8].

Given a polynomial pair (f,g), we want to find many coprime pairs such that $N(a-b\alpha)$ and $a-bm$ are both smooth with respect to some integers B_f, B_g, where B_f is algebraic boundary, and B_g is the rational boundary. In practice, the notations $F(x,y)$ and $G(x,y)$ for the homogenized polynomials corresponding to $f(x)$ and $g(x)$ are often used.

Step 3: solving equations [9][10].

Solving equations can be divided into structured Gaussian elimination, solving equations. Structured Gaussian elimination contains: removing duplicates, discarding singletons and so on. We used Block Lanczos algorithm and Block Wiedemann algorithm to solve the huge sparse equations over the field $\mathbb{F}_2$.

Step 4: finding square roots [11][12].

In the final step, we want to find the square roots $\prod_{(a,b)\in S}(a-b\alpha)$. The algorithms for finding square roots consist of UFD method, method of Couveignes, Montgomery-Nguyen, and Emmanuel Thome methods. They are based on the Chinese Remainder Theorem.

Through the above four steps, according to the homomorphism mapping, get $(x,y)\in\mathbb{Z}^2$, satisfy $x^2\equiv y^2 \pmod{n}$, which may give a factor of n with probability at least $\frac{1}{2}$.

The space complexity analysis of sieving have important significance for the number field sieve.

$p-$value of every root

If $p\nmid\Delta_h$, the evaluation of a root $\xi\in\mathbb{P}^1(\mathbb{Z}/p^e\mathbb{Z})$ is given in[7]:

$$p_r\left(\upsilon_p\left(h(X)\right)=e\mid X=\xi\right)=\frac{1}{p^e+p^{e-1}}\times\frac{p^e-p^{e-1}}{p^e} \tag{1}$$

where $X\in\mathbb{Z}$ is a random and uniformly distributed in probability space $(\Omega,\Im,p_r)$, $\upsilon_p\left(h(X)\right)$ denotes the exponent of the largest power of p dividing the integer $h(X)$, p^e-p^{e-1} is the number of the elements on each congruence class.

On the basis of Shi Bai' work, we give the simply former of (1) as follows:

$$p_r\left(\upsilon_p\left(h(X)\right)=e\mid X=\xi\right)=\frac{1}{p^e+p^{e-1}}\times\frac{p-1}{p} \tag{2}$$

Expected $p-$value provided by every root on the finite field

By (2) we can get:

Lemma 1 Let $X,Y\in\mathbb{Z}$ be independent random variables uniformly distributed in a probability space $(\Omega,\Im,p_r)$ and satisfy $\gcd(X,Y)=1$, if $H(x,y)\bmod p^1=0$ has a single root, then expected $p-$value provided by every single root is: $E\left(\upsilon_p\left(H(X,Y)\right)\right)=\frac{p}{p^2-1}$.

Prove By (2), expected $p-$value provided by every single root is:

$$E\left(\upsilon_p\left(H(X,Y)\right)\right)=\sum_{i=0}^{\infty}\left(i\times p_r\left(v_p(h)=i\right)\right)$$

$$=\sum_{e=1}^{\infty}e\frac{1}{p^{e-1}(p+1)}\left(\frac{p-1}{p}\right)=\frac{p-1}{p}\frac{1}{p+1}\frac{p^2}{(p-1)^2}=\frac{1}{p+1}\frac{p}{(p-1)}=\frac{p}{p^2-1}.$$

For ramified prime [13], and the empirical formula is given by Shi Bai as follows:

$$E\left(\upsilon_p\left(H(X,Y)\right)\right)=\frac{1}{p+1}\sum_{e=1}^{d}\frac{n_{p,e}}{p^{e-1}}$$

This paper gives the analysis of multiple root for unramified prime on the finite field, the average $p-$valuation is given.

Lemma 2 Let $X,Y\in\mathbb{Z}$ be independent random variables uniformly distributed in a probability space $(\Omega,\Im,p_r)$ and satisfy $\gcd(X,Y)=1$, if $H(x,y)\bmod p^1=0$ has an multiple root, then expected $p-$value provided by every multiple root is: $E\left(\upsilon_p\left(H(X,Y)\right)\right)=\frac{3\times p^2-3p+1}{(p-1)p(p+1)}$.

Prove When $\xi\bmod p$ is a multiple root of $H(X,Y)$, in the process of root lifting, according to

the root lifting, it will be terminated or keep p root lifting, so the expected $p-$value is:

$$E\left(\upsilon_p\left(H\left(X,Y\right)\right)\right)$$

$$=\frac{1}{p+1}\left(1\times\frac{p-1}{p}+2\times\left(p\right)\times\frac{1}{p}\times\frac{p-1}{p}+3\times\left(p\right)\times\frac{1}{p^2}\times\frac{p-1}{p}+\cdots\right)$$

$$=\frac{p-1}{p}\times\frac{1}{p+1}\left(1+2+3\times\frac{1}{p}+4\times\frac{1}{p^2}+\cdots\right)$$

$$=\frac{p-1}{p}\times\frac{3}{p+1}+\frac{1}{p}\times\frac{3}{p+1}+\frac{1}{p}\times\frac{1}{p+1}\times\frac{1}{p-1}=\frac{3\times\left(p-1\right)^2+3\times\left(p-1\right)+1}{\left(p-1\right)p\left(p+1\right)}=\frac{3\times p^2-3p+1}{\left(p-1\right)p\left(p+1\right)}.$$

Polynomial selection

Polynomial selection can be divided into four steps: polynomial generation, size optimization, root optimization and sieving test. Root optimization gives two optimal objective function, namely α function and Murphy$E$ function.

Let $Z\in\square$ be independent random variables uniformly distributed in a probability space $\left(\Omega,\Im,p_r\right)$, for polynomial h,α function is defined as:

$$\alpha\left(H\right)=\sum_{p\le B_h}\left(E\left(\nu_p\left(Z\right)\right)-E\left(\nu_p\left(H\left(X,Y\right)\right)\right)\right)\log p$$

By using the method in this paper, we give a kind of proof for $E\left(\nu_p\left(\square\right)\right)$ as follows:

Lemma 3 $E\left(\nu_p\left(Z\right)\right)=\frac{1}{p-1}$.

Prove

$$E\left(\nu_p\left(Z\right)\right)$$

$$=p_r\left(p\mid Z\right)\sum_{e=1}^{\infty}e\left(p^{e+1}\nmid Z\mid p^e\mid Z\right)=\frac{1}{p}\times\left(\frac{p-1}{p}+2\times\frac{1}{p}\times\frac{p-1}{p}+\cdots\right)$$

$$=\frac{1}{p}\times\frac{p-1}{p}\times\frac{1}{\left(1-\frac{1}{p}\right)^2}=\frac{1}{p}\times\frac{p-1}{p}\times\frac{p^2}{\left(p-1\right)^2}=\frac{1}{p-1}.$$

If p is the unramified prime, let n_p be the number of single roots of $H\left(X,Y\right)\bmod p=0$, then by lemma 1, $E\left(\nu_p\left(H\left(X,Y\right)\right)\right)=\frac{n_p p}{p^2-1}$.

If p is the ramified prime, let the roots of $H\left(X,Y\right)\bmod p=0$ be: $n_{p,0,1},n_{p,0,2},\cdots,n_{p,0,\varepsilon_p^s},n_{p,1,1},n_{p,1,2},\cdots,n_{p,1,\varepsilon_p^m}$, where $n_{p,0,1},n_{p,0,2},\cdots,n_{p,0,\varepsilon_p^s}$ are single roots, $n_{p,1,1},n_{p,1,2},\cdots,n_{p,1,\varepsilon_p^m}$ are different multiple roots, then according to lemma 1 and lemma 2, we can get:

$$E\left(\nu_p\left(H\left(X,Y\right)\right)\right)=\frac{\varepsilon_p^s p}{p^2-1}+\frac{\varepsilon_p^m\left(3\times p^2-3p+1\right)}{\left(p-1\right)p\left(p+1\right)}.$$

So we can get the following theorem:

Theorem 4 If p is unramified prime: $E\left(\nu_p\left(H\left(X,Y\right)\right)\right)=\frac{n_p p}{p^2-1}$.

If p is ramified prime: $E\left(\nu_p\left(H\left(X,Y\right)\right)\right)=\frac{\varepsilon_p^s p}{p^2-1}+\frac{\varepsilon_p^m\left(3\times p^2-3p+1\right)}{\left(p-1\right)p\left(p+1\right)}$.

Space complexity of sieving

Lemma 5 In each smooth pairs, the amount of data which have p factor in rational side on average is $\frac{1}{p+1}$.

Prove

$$\sum_{i=1}^{\infty}\frac{1}{p^{i-1}(p+1)}\left(\frac{p-1}{p}\right)$$

$$=1\times\frac{1}{p+1}\times\frac{p-1}{p}+1\times\frac{1}{p(p+1)}\times\frac{p-1}{p}+1\times\frac{1}{p^2(p+1)}\times\frac{p-1}{p}+\cdots$$

$$=\frac{1}{p+1}\times\frac{p-1}{p}\times\left(1+\frac{1}{p}+\frac{1}{p^2}+\frac{1}{p^3}+\cdots\right)=\frac{1}{p+1}\times\frac{p-1}{p}\times\frac{1}{1-1/p}=\frac{1}{p+1}.$$

By lemma 5, we can get:

Theorem 6 If the amount of smooth pairs is Γ, then space complexity of data in rational side is: $C_{s,g}^{s}=\Gamma\times\left(\ln\ln\left(B_g\right)^2+O(1)\right)$.

Prove

$$C_{s,g}^{s}=\Gamma\times\sum_{p\le B_g}\frac{1}{p+1}=\Gamma\times\sum_{p\le B_g}\frac{1}{p}\frac{1}{1+\frac{1}{p}}=\Gamma\times\sum_{p\le B_g}\frac{1}{p}\left(1-\frac{1}{p}+\frac{1}{p^2}-\frac{1}{p^3}+\cdots\right)$$

$$=\Gamma\times\left(\sum_{p\le B_g}\frac{1}{p}+\sum_{p\le B_g}\left(\frac{1}{p^3}+\frac{1}{p^5}+\cdots\right)-\sum_{p\le B_g}\left(\frac{1}{p^2}+\frac{1}{p^4}+\frac{1}{p^6}+\cdots\right)\right)$$

$$=\Gamma\times\left(\ln\ln\left(B_g\right)^2+O(1)\right).$$

where $C_{s,g}^{s}$ denotes the rational space complexity of sieving.

By lemma 5 we can easily get:

Lemma 7 If p is unramified prime, assuming $H(x,y)\bmod p=0$ have ε_p single roots, then in each smooth pairs, the amount of data which have p factor in algebraic side on average is $\frac{\varepsilon_p}{p+1}$.

By the theorem 5 and lemma 2 we can get:

Lemma 8 If p is a ramified prim , assuming $H(x,y)\bmod p=0$ have ε_p^s single roots and ε_p^m multiple roots, then space complexity of data in algebraic side is $\frac{\varepsilon_p^s p+\varepsilon_p^m(2p-1)}{p(p+1)}$.

By theorem 7 and lemma 8, we can get:

Lemma 9 Let S_r be a set of ramified prime, the amount of data in a smooth pair in algebraic side is: $\sum_{p\le B_f,\,p\notin S_r}\frac{\varepsilon_p}{p+1}+\sum_{p\le B_f,\,p\in S_r}\frac{\varepsilon_p^s p+\varepsilon_p^m(2p-1)}{p(p+1)}$.

By the lemma 9, we can get:

Lemma 10 The amount of data in a smooth pair in algebraic side at most $(d+1)\times\left(\ln\ln\left(B_f\right)^2+O(1)\right)$.

Prove

$$\sum_{p\le B_f,\, p\notin S_r}\frac{\varepsilon_p}{p+1}+\sum_{p\le B_f,\, p\in S_r}\frac{\varepsilon_p^s p+\varepsilon_p^m(2p-1)}{p(p+1)}$$

$$\le\sum_{p\le B_f,\, p\notin S_r}\frac{d+1}{p+1}+\sum_{p\le B_f,\, p\in S_r}\frac{(\varepsilon_p^s+2\varepsilon_p^m)p-\varepsilon_p^m}{p(p+1)}\le\sum_{p\le B_f,\, p\notin S_r}\frac{d+1}{p+1}+\sum_{p\le B_f,\, p\in S_r}\frac{(d+1)p}{p(p+1)}$$

$$=\sum_{p\le B_f}\frac{d+1}{p+1}=(d+1)\times\sum_{p\le B_f}\frac{1}{p+1}\le(d+1)\times\sum_{p\le B_f}\frac{1}{p}=(d+1)\times\left(\ln\ln(B_f)^2+O(1)\right).$$

where $C_{s,f}^s$ represents the algebraic space complexity of sieving, $d=\deg(f)$.

By lemma 10, we can get:

Theorem 11 If the amount of smooth pairs is Γ, then space complexity of data in algebraic side is: $C_{s,f}^s\le\Gamma\times(d+1)\times\left(\ln\ln(B_f)^2+O(1)\right)$.

By theorem 6 and theorem 11, we can get:

Inference 12 If the amount of smooth pairs is Γ, then the total space complexity of sieving is:

$$C_s^s=C_{s,g}^s+C_{s,f}^s\le\Gamma\times\left(\ln\ln(B_g)^2+O(1)\right)+\Gamma\times(d+1)\times\left(\ln\ln(B_f)^2+O(1)\right)$$

$$C_s^s=C_{s,g}^s+C_{s,f}^s\ge\Gamma\times\left(\left(\ln\ln(B_g)^2+\ln\ln(B_f)^2\right)+O(1)\right).$$

Inference 13 If the amount of smooth pairs is Γ and $B_g=B_f=B$, then the total space complexity is:

$$C_s^s=C_{s,g}^s+C_{s,f}^s\le\left(\Gamma\times\left(\ln\ln(B)^2+O(1)\right)\right)(2+d).$$

$$C_s^s=C_{s,g}^s+C_{s,f}^s\ge 2\Gamma\times\left(\left(\ln\ln(B)^2+O(1)\right)+O(1)\right).$$

Test results

(1)Assume X,Y be random variable which independent and uniformly distributed in a probability space $(\Omega,\Im,p_r)$, where $\gcd(X,Y)=1$, if $H(x,y)\bmod p^1=0$ has multiple roots, where $H(x,y)$ be the homogenized polynomials corresponding to the irreducible polynomial $h(x)$ on □, assuming X,Y leading to the values $H(X,Y)$ is uniformly random and $H(X,Y)\bmod p^e$ $(\forall e\ge 1)$ uniformly distributed, and it is provide each multiple root expected $p-$value as follows: $E\left(\upsilon_p\left(H(X,Y)\right)\right)=\frac{3\times p^2-3p+1}{(p-1)p(p+1)}$,where $\upsilon_p\left(H(X,Y)\right)$ denotes the exponent of the largest power of p dividing integer $H(X,Y)$.

(2) We give the space complexity analysis results for sieving:

$$C_s^s=C_{s,g}^s+C_{s,f}^s\le\Gamma\times\left(\ln\ln(B_g)^2+O(1)\right)+\Gamma\times(d+1)\times\left(\ln\ln(B_f)^2+O(1)\right),$$

$$C_s^s=C_{s,g}^s+C_{s,f}^s\ge\Gamma\times\left(\left(\ln\ln(B_g)^2+\ln\ln(B_f)^2\right)+O(1)\right)$$

where C_s^s defines the space complexity of sieving.

Based on the above results, we can further estimate the space complexity and the computational complexity for structured Gaussian elimination and solving equations in the number sieve integer factorization.

The results in this paper have important significance for solving discrete logarithm in the number field sieve.

Conclusion

In this paper, the $p-adic$ evaluation provided by each root is analyzed. The space complexity of sieving over the ring $Z/2$ is also presented. The implementation can be found in CADO-NFS[14].

Acknowledgement

This work is supported by the State Key Basic Research and Development Plan (2013CB338003), and the Milky Way High Performance Computing foundation.

References

[1] W. Diffie , M. Hellman, New direction in cryptography, IEEE Transations in Informayion Theory.22 (1976),644-654.

[2] R. L. Rivest, A. Shamir, H. Adleman, A method for obtaining digital signatures and public-keycryptosystems, Communication of the ACM,21(2) 1978,120-126.

[3] A. K. Lenstra, H. W. Lenstra, Jr., editors, The Development of the Number Field Sieve, volume 1554 of Lecture Notes in Mathematics. Springer,1993.

[4] T. Kleinjung, K. Aoki, J. Franke, A. K. Lenstra, E. Thome, J. W. Bos, P. Gaudry, A. Kruppa, P.L. Montgomery, D.A. Osvik , H.J.J. te Riele, A. Timofeev, P. Zimmermann, Factorization of a 768-bit RSA modulus .In Proceedings of CRYPTO '10,Lecture Notes in Computer Science,6223(2010), 333-350, Springer.

[5] B. A. Murphy, Polynomial selection for the number field sieve integer factorization algorithm, PhD thesis, The Australian National University,1999.

[6] T.Kleinjung, On polynomial selection for the general number field sieve,Mathematics of Computation,75(256):2037-2047, 2006.

[7] Shi Bai, Polynomial Selection for the Number Field Sieve,PhD thesis,The Australian National University. 2011.

[8] J. M. Pollard, The lattice sieve, Lecture Notes in Mathematics, Springer-Verlag, 43-49,1554.

[9] D. H. Wiedemann, Sloving sparse linear equation equations over finite field, IEEE T rans. Inform. Theory 32(1986),54-62.

[10] D. Coppersmith, Solving linear equation over GF(2) via block Wiedemann algorithm, Mathematics of Computation62,05(1994),333—350.

[11] P. L .Montgomery, A Block Lanczos algorithm for finding dependencies over GF(2),Lecture Notes in Computer Science.921:106-120,1995.

[12] P. L. Montgomery, Square roots of products of algebraic numbers (1997),unpublished draft, significantly different from published version[14] (May 16, 1997).

[13] I.M.Gelfand, M.M.Kapranov,A.V.Zelevinsky,Discriminants,Resultants,and Multidimensional Determinants.The world book publishing company.

[14] http://cado-nfs.gforge.inria.fr,2011.