Estimation of Cancellation Errors in Multivariate Hensel Construction with Floating-Point Numbers

Kosaku Nagasaka
Doctoral Program of Mathematics, Univ. of Tsukuba.
nagasaka@math.tsukuba.ac.jp

Abstract

Multivariate Hensel construction with floating-point numbers often cause large cancellation errors which are errors due to a cancellation of almost the same numbers. Sasaki and Yamaguchi [SY98] showed that multivariate Hensel construction causes large cancellation errors if the expansion point is chosen near a singular point, and Sasaki [Sas00] studied four mechanisms of term cancellations near a singular point. However, an analysis of cancellation errors, if the expansion point is chosen randomly, has never been studied. Moreover, in practical computations, we do not know in advance where the nearest singular point is. In this paper, we investigate a distance to the nearest singular point from the expansion point and estimate a magnitude of cancellation errors.

1 Introduction and Notations

Multivariate Hensel construction is used for many algebraic computations such as multivariate polynomial factorization, power-series expansions and so on. Recently, many researchers are interested in algebraic computations over approximate arithmetic domains, such as floating-point numbers. Using floating-point numbers, each arithmetic operations may cause round-off errors or cancellation errors. Especially, cancellation errors are critical. For example, the following operation causes a large cancellation error (9 significant digits are lost).

$$0.\underline{3674301083421}?? - 0.\underline{36743010823 89}?? = \underline{1.032}???????????? \times 10^{-10}, \tag{1.1}$$

where underlined figures are significant digits. Hence, we must consider cancellation errors to handle the multivariate Hensel construction with floating-point numbers.

Although the multivariate Hensel construction is widely used, its numerical analysis is not given without the following two papers. Sasaki and Yamaguchi [SY98] showed that multivariate Hensel construction causes large cancellation errors if the expansion point is chosen near a singular point. Their analysis is based on Cauchy-Hadamard's theorem and cancellation errors

are occurred for the crossing case (factors are integral power-series at a singular point). The following four mechanisms of term cancellations are studied by Sasaki [Sas00]. The first two mechanisms are exact and approximate cancellations based on a property of Moses-Yun's interpolation polynomials. The third and fourth mechanisms are based on the extended Hensel construction ([SK99] and [SI00]).

However, for practical computations, they are not applicable in the following two points.

- Usually, we do not know where the nearest singular point is.
- Usually, we do not know whether factors are crossing or not.

Therefore, in these points, we investigate how large cancellation errors are occurred. In **2**, we prepare the multivariate Hensel construction. Estimating the nearest singular point and cancellation errors are discussed in **3** and **4**, respectively. We give numerical examples in **5**, and conclusion in **6**.

We use the following notations.

$F(x, u_1, \ldots, u_\ell)$	:	the given multivariate polynomial in $\mathbb{C}[x, u_1, \ldots, u_\ell]$;
$x, u_1, \ldots, u_\ell$	:	variables, x is the main variable;
$\boldsymbol{u}$	:	an abbreviation for $u_1, \ldots, u_\ell$ (e.g. $F(x, \boldsymbol{u}) = F(x, u_1, \ldots, u_\ell)$);
$\deg(P)$	:	a degree w.r.t. x , of a polynomial P , and we put $n = \deg(F)$;
$\omega_1, \ldots, \omega_n$	:	the roots of $F(x, 0)$;
S	:	$\langle u_1, \ldots, u_\ell \rangle$, a polynomial ideal generated by $u_1, \ldots, u_\ell$;
$\mathrm{Syl}(P_1, P_2)$	:	the Sylvester matrix for polynomials P_1 and P_2;
$\|P\|_p$	:	the polynomial norms of a polynomial $P = \sum_{i=0}^{n} c_i x^i$, $\|P\|_p = (\sum_i \|c_i\|^p)^{1/p}$ $(p < \infty)$, $\|P\|_\infty = \max_i \|c_i\|$;
$\gcd(P_1, P_2)$	:	the greatest common divisor of polynomials P_1 and P_2.

We assume that $F(x, \boldsymbol{u})$ is monic and square-free w.r.t. x and $F_1(x, \boldsymbol{u}) \neq 0$ where

$$\begin{gathered} F(x, \boldsymbol{u}) = F_0(x, \boldsymbol{u}) + F_1(x, \boldsymbol{u}) + \cdots + F_{e-1}(x, \boldsymbol{u}) + F_e(x, \boldsymbol{u}), \\ e = \deg_{u_1, \ldots, u_\ell}(F), \quad \deg_{u_1, \ldots, u_\ell}(F_i) = i \ \ (i = 0, 1, \ldots, e). \end{gathered} \tag{1.2}$$

Let the expansion point of Hensel factors be the origin. This means that we apply a suitable variable transformation if the origin is a singular point of $F(x, \boldsymbol{u})$. Furthermore, without loss of generality, we assume that $F(x, 0)$ is monic and square-free w.r.t. x. We investigate the construction under the following restriction **R**.

$$\mathbf{R}\text{: } F(x, \boldsymbol{u}) \text{ is not sparse, } \|F(x, \boldsymbol{u})\|_\infty \simeq 1 \text{ and } \|F_1(x, \boldsymbol{u})\|_1 \gg \|F_j(x, \boldsymbol{u})\|_1 \ \ (j > 1). \tag{1.3}$$

For monitoring cancellation errors due to the Hensel construction, we use *effective floating-point number* which is proposed by Kako and Sasaki [KS97]. Although *Mathematica*[1] has similar numbers, it is not applicable since it handled round-off errors together.

[1] Wolfram Research Inc.

2 Multivariate Hensel Construction

Let $G^{(0)}$ and $H^{(0)}$ be initial factors of the multivariate Hensel construction as follows.

$$G^{(0)} = \prod_{i \in N_G} (x - \omega_i), \quad H^{(0)} = \prod_{i \in N_H} (x - \omega_i), \quad m = \#N_G, \tag{2.1}$$

where N_G and N_H are the set of integers and satisfying $N_G \cap N_H = \phi$ and $N_G \cup N_H = \{1, \ldots, n\}$.

The multivariate Hensel construction is to calculate polynomials $G^{(k)}(x, \boldsymbol{u})$ and $H^{(k)}(x, \boldsymbol{u})$, $k = 1, 2, \ldots$, satisfying

$$F(x, \boldsymbol{u}) \equiv G^{(k)}(x, \boldsymbol{u}) H^{(k)}(x, \boldsymbol{u}) \pmod{S^{k+1}}. \tag{2.2}$$

The construction method is as follows. First, using the extended Euclidean algorithm, we calculate Moses-Yun's interpolation polynomials $A_i(x)$ and $B_i(x)$ $(i = 0, \ldots, n-1)$ which satisfy

$$\begin{cases} A_i(x) H^{(0)}(x) + B_i(x) G^{(0)} = x^i, \\ \deg(A_i) < \deg(G^{(0)}), \ \ \deg(B_i) \leq \deg(H^{(0)}). \end{cases} \tag{2.3}$$

Suppose that we have constructed $G^{(\kappa)}$ and $H^{(\kappa)}$, $\kappa = 0, 1, \ldots, k-1$. We calculate

$$\text{Step 1} \qquad \underline{D}^{(k)} \equiv G^{(k-1)} H^{(k-1)} \pmod{S^{k+1}}. \tag{2.4}$$

$$\text{Step 2} \qquad D^{(k)} \equiv F - \underline{D}^{(k)} \equiv \sum_{i=0}^{n-1} d_i^{(k)} x^i \pmod{S^{k+1}}. \tag{2.5}$$

$$\text{Step 3} \qquad G_k = \sum_{i=0}^{n-1} A_i d_i^{(k)}, \quad H_k = \sum_{i=0}^{n-1} B_i d_i^{(k)}. \tag{2.6}$$

Then, we construct $G^{(k)}$ and $H^{(k)}$ as

$$G^{(k)} = G^{(k-1)} + G_k, \quad H^{(k)} = H^{(k-1)} + H_k. \tag{2.7}$$

3 Distance from Singular Point

In this section, we investigate a distance to the nearest singular point from the origin.

Definition 1 *We denote δ as the minimum Euclidean distance from the origin to a singular point of $F(x, \boldsymbol{u})$. Therefore, at that point $\boldsymbol{v}$, $F(x, \boldsymbol{v})$ is not square-free w.r.t. x.* $\triangleleft$

Let $\omega_1^*, \ldots, \omega_n^*$ be the roots of $F(x, \boldsymbol{v})$, minimizing $\sum_{i=1}^n |\, \omega_i - \omega_i^* |$, and put

$$\Delta_G = G^{(0)} - G^{(0)^*}, \quad \Delta_H = H^{(0)} - H^{(0)^*}, \quad G^{(0)^*} = \prod_{i \in N_G} (x - \omega_i^*), \quad H^{(0)^*} = \prod_{i \in N_H} (x - \omega_i^*).$$

By the restriction **R**, the case $\delta \gg 1$ is not occurred in most cases. Moreover, [Sas00] says, under their restrictions, large cancellation errors are not occurred if the expansion point is located far from the nearest singular point. Hence, we consider the cases $\delta \simeq 1$ and $\delta \ll 1$ only. Then, we have

$$\|F_1(x, \boldsymbol{v})\|_1 \simeq \|G^{(0)}\|_1\|\Delta_H\|_1 + \|H^{(0)}\|_1\|\Delta_G\|_1 + \|\Delta_G\|_1\|\Delta_H\|_1, \tag{3.1}$$

by the following equalities.

$$\begin{aligned} F(x,0) - F(x,\boldsymbol{v}) &= G^{(0)}H^{(0)} - G^{(0)^*}H^{(0)^*} \\ &= G^{(0)}H^{(0)} - (G^{(0)} - \Delta_G)(H^{(0)} - \Delta_H) \\ &= G^{(0)}\Delta_H + \Delta_G H^{(0)} - \Delta_G\Delta_H \\ &= F_1(x,\boldsymbol{v}) + \cdots + F_e(x,\boldsymbol{v}). \end{aligned} \tag{3.2}$$

Calculating $G^{(0)^*}$ and $H^{(0)^*}$ is a kind of the approximate GCD problem. We explain the approximate GCD briefly, as in the work by Beckermann and Labahn [BG98].

Definition 2 ([BG98] ϵ-prime) *Let*

$$\epsilon(G^{(0)}, H^{(0)}) := \inf \max\{\|G^{(0)} - G^*\|_1, \|H^{(0)} - H^*\|_1\}, \tag{3.3}$$

where G^* and H^* have a common root, $\deg(G^*) \leq m$ and $\deg(H^*) \leq n-m$. We will then refer to $G^{(0)}, H^{(0)}$ as being ϵ-prime. $\gcd(G^*, H^*)$ is an approximate GCD of $G^{(0)}$ and $H^{(0)}$. ◁

Lemma 1 ([BG98] Lemma 2.1) *We have*

$$\epsilon(G^{(0)}, H^{(0)}) \geq \frac{1}{\|\mathrm{Syl}(G^{(0)}, H^{(0)})^{-1}\|_1}. \tag{3.4}$$

◁

Lemma 2 ([BG98] Corollary 3.2) *Let*

$$\kappa := \max\{\|A_0 + x^m B_0\|_1, \|A_{n-1} + x^m B_{n-1}\|_1\}. \tag{3.5}$$

We have

$$\kappa \leq \|\mathrm{Syl}(G^{(0)}, H^{(0)})^{-1}\|_1 \leq \kappa + 2\,\|B_{n-1}A_0 - A_{n-1}B_0\|_1 \max\{\|G^{(0)}\|_1, \|H^{(0)}\|_1\}. \tag{3.6}$$

◁

$\gcd(G^{(0)^*}, H^{(0)^*})$ may be not an approximate GCD of $G^{(0)}$ and $H^{(0)}$ since possible their variations are constrained. Thus we have

$$\epsilon(G^{(0)}, H^{(0)}) \leq \max\{\|G^{(0)} - G^{(0)^*}\|_1, \|H^{(0)} - H^{(0)^*}\|_1\} = \max\{\|\Delta_G\|_1, \|\Delta_H\|_1\}. \tag{3.7}$$

Theorem 1 *By the above discussions, we estimate δ as*

$$\delta = (\kappa^{-1}\,\|G^{(0)}\|_1 + \kappa^{-1}\,\|H^{(0)}\|_1 + \kappa^{-2}) / \,\|F(x, \boldsymbol{u})\|_1\,. \tag{3.8}$$

◁

We note that calculating κ is done by (2.3) which is needed to calculate for Hensel construction. Hence, unnecessary computations for the construction are not needed to calculate the above δ.

4 Estimation of Cancellation Errors

Suppose that we attempt to compute $P(x, \boldsymbol{u}) = G(x, \boldsymbol{u})H(x, \boldsymbol{u})$. If $\| G(x, \boldsymbol{u}) \|_\infty$ and $\| H(x, \boldsymbol{u}) \|_\infty$ are larger than $\| P(x, \boldsymbol{u}) \|_\infty$, then term cancellations are occurred. In this section, to estimate cancellation errors, we estimate $\| G^{(k)} \|_\infty$, $\| H^{(k)} \|_\infty$ and their multipliers norms.

As in [SY98], we define the concepts *branching* and *crossing*.

Definition 3 ([SY98] branching and crossing) *Let t be the total-degree variable for $u_1, \ldots, u_\ell$, $F(x, t\boldsymbol{u})$ be a multivariate polynomial which has a singular point at the origin and be factorized as*

$$F(x, t\boldsymbol{u}) = G(x, t\boldsymbol{u})H(x, t\boldsymbol{u}), \tag{4.1}$$

where $G(x, t\boldsymbol{u})$ and $H(x, t\boldsymbol{u})$ are polynomials in x and fractional power series in t. Note that the fractional power series includes the only integral power series.

If $G(x, t\boldsymbol{u})$ and $H(x, t\boldsymbol{u})$ are only integral power series in t, then they are crossing at the origin. Otherwise, they are branching. ◁

By the Cauchy-Hadamard's theorem, we see for $k \gg 1$ that

$$\begin{aligned} &\frac{\|G^{(k+1)}\|_\infty}{\|G^{(k)}\|_\infty}, \frac{\|H^{(k+1)}\|_\infty}{\|H^{(k)}\|_\infty} = O(\delta^{-1}) \quad \text{in branching case,} \\ &\frac{\|G^{(k+1)}\|_\infty}{\|G^{(k)}\|_\infty}, \frac{\|H^{(k+1)}\|_\infty}{\|H^{(k)}\|_\infty} = O(\delta^{0}) \quad \text{in crossing case,} \end{aligned} \tag{4.2}$$

where O denotes Landau's order symbol. Hence, in the branching cases, we have

$$\|G_k\|_\infty \approx \|G_1\|_\infty + (k-1)\delta^{-1}, \quad \|H_k\|_\infty \approx \|H_1\|_\infty + (k-1)\delta^{-1}. \tag{4.3}$$

We consider ∞-norms of G_k and H_k.

$$\|G_k\|_\infty = \|\sum_{i=0}^{n-1} A_i d_i^{(k)}\|_\infty \approx \max_i \|A_i\|_\infty \mid d_i^{(k)}|. \tag{4.4}$$

$$\|H_k\|_\infty = \|\sum_{i=0}^{n-1} B_i d_i^{(k)}\|_\infty \approx \max_i \|B_i\|_\infty \mid d_i^{(k)}|. \tag{4.5}$$

By (2.4), we have $\|\underline{D}^{(k)}\|_\infty \leq \max_{i+j=k, i,j<k} \|G_i\|_\infty \|H_j\|_\infty = \|G_1\|_\infty + \|H_1\|_\infty + (k-1)\delta^{-1}$. Therefore, the difference between (4.3), (4.4) and (4.5) may cause cancellation errors. There is

not only the case that the dominant terms of A_i and $D^{(k)}$ are multiplied but also not multiplied for the computations of G_k (and H_k). To handle both cases, we use the following simple model.

$$\max_i \|A_i\|_\infty \mid d_i^{(k)}| \approx \|A_{j_1}\|_\infty \|\underline{D}^{(k)}\|_\infty, \tag{4.6}$$

$$\max_i \|B_i\|_\infty \mid d_i^{(k)}| \approx \|B_{j_2}\|_\infty \|\underline{D}^{(k)}\|_\infty, \tag{4.7}$$

where j_1 and j_2 maximizes the followings, respectively.

$$\begin{array}{l} A_G = \max_{j_1} \|A_{j_1}\|_\infty \mid \text{coefficients of } x^{j_1}, \text{ of } G_1(x, \boldsymbol{u})H_1(x, \boldsymbol{u})|, \\ B_H = \max_{j_2} \|B_{j_2}\|_\infty \mid \text{coefficients of } x^{j_2}, \text{ of } G_1(x, \boldsymbol{u})H_1(x, \boldsymbol{u})|. \end{array} \tag{4.8}$$

Theorem 2 *In the branching cases, we estimate a magnitude of cancellation errors as*

$$k \times \max\{A_G/(\|G_1(x, \boldsymbol{u})\|_\infty \times\delta),\ B_H/(\|G_1(x, \boldsymbol{u})\|_\infty \times\delta)\}. \tag{4.9}$$

◁

Remark 1 For the crossing cases, we have to investigate other estimation. However, as in the following section, our estimation is enough to use for randomly chosen polynomials. ◁

5 Numerical Examples

We have tested our estimation formulas as follows, using the double precision floating-point numbers. We have generated 100 bivariate polynomials of degrees n and e w.r.t. x and u_1, respectively, with coefficients randomly chosen in the range $[-1, 1]$. We have constructed Hensel factors up to degree k. Figure 1 and 2 shows the results, where the "Actual δ" denotes the mean of $\|G^{(k+1)}\|_\infty$ / $\|G^{(k)}\|_\infty$ and $\|H^{(k+1)}\|_\infty$ / $\|H^{(k)}\|_\infty$, the "Estimation of δ" denotes the our estimation in (3.8) and the "reducible" and "irreducible" denote each sample is the product of two distinct irreducible polynomials or not, respectively (however, initial factors are taken all possible combinations among the roots of $F(x, 0)$). Figure 3 and 4 shows the results, where the "Actual Errors" denotes the magnitude of cancellation errors occured in computations and the "Estimation of Cancellation Errors" denotes the our estimation discussed in the previous section, using actual δs instead of (3.8). All figures are the common logarithm plots and indeces are corresponding to the number of digits.

In Fig. 1, we see that our estimations of δ are close to the actual δ. On the other hand, in Fig. 2, δ are over-estimated. The author thinks that these over-estimations are caused by the reducibilities of given polynomials, and the differences between both sides of inequality (3.7) may be large. In fact, if Hensel factors $G^{(k)}$ and $H^{(k)}$ are exactly or approximately polynomial factors of F, then they may be crossing and their maximum coefficients may be not large.

In Fig. 3, we see that irreducible polynomials do not cause large cancellation errors in this example. On the other hand, in Fig. 4, almost cancellation errors are estimated. Some date located in the upper side are the cases in which Hensel factors are polynomial factors of the given polynomials, and they are not critical. Because, we have $D^{(k)} \simeq 0$ by the reducibilities of polynomials, and it means all significant digits are lost.

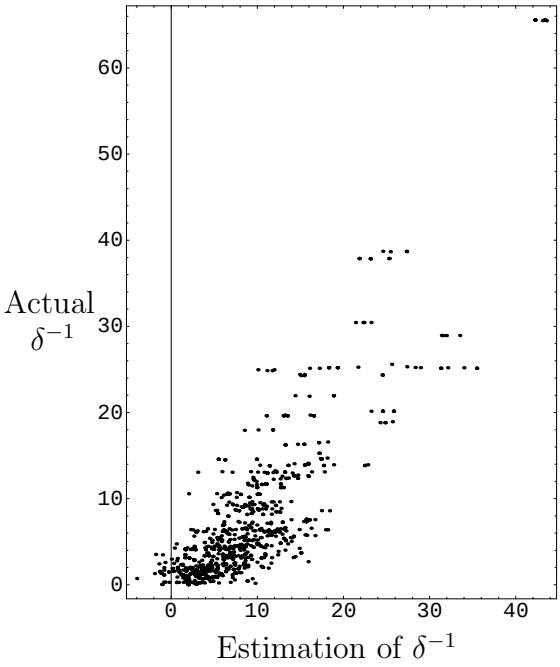

Fig. 1: irreducible, $n = 5$, $e = 5$ and $k = 24$

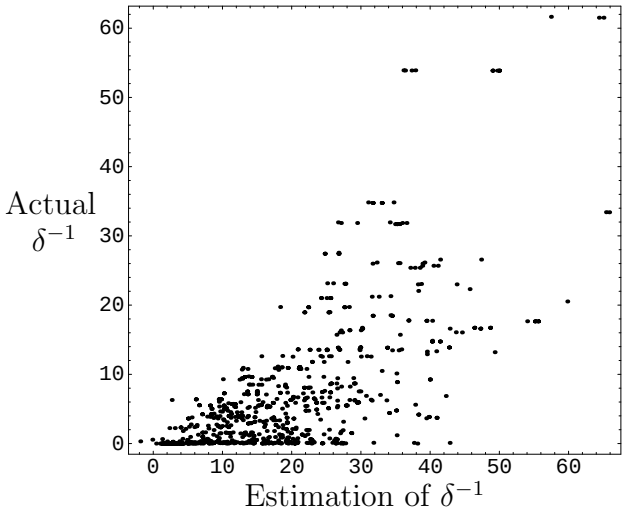

Fig. 2: reducible, $n = 5$, $e = 5$ and $k = 24$

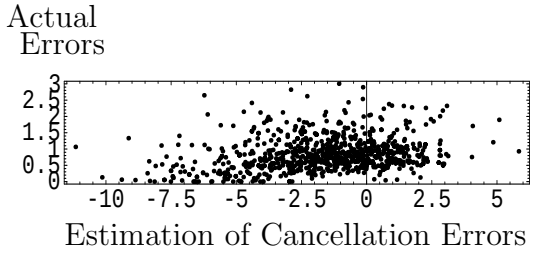

Fig. 3: irreducible, $n = 9$, $e = 11$ and $k = 12$

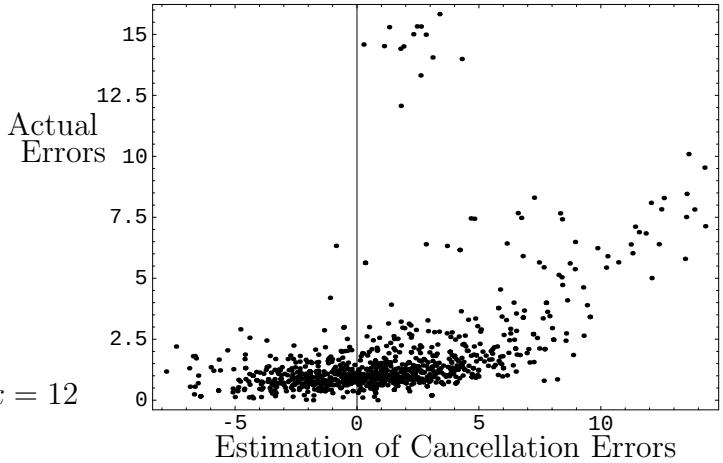

Fig. 4: reducible, $n = 9$, $e = 11$ and $k = 12$

6 Conclusion

We investigate the cancellation errors in the multivariate Hensel construction. The numerical examples show that these estimation are able to use for practical computations if we sample an actual δ. Our estimation in (4.9) does not detect the small cancellation errors as in Fig. 3, however, it does for the large cancellation errors as in Fig. 4. We use the big precision numbers if the large errors are estimated.

However, if we do not sample an actual δ, then the estimation in (3.8) may be over-estimated for reducible polynomials and the estimation of cancellation errors is broken (hence, in Fig. 4, we use the actual δ instead of our estimations). More efficient δ estimation is needed to estimate the cancellation errors in the reducible case.

References

[BG98] B. Beckermann and G. Labahn. When are two numerical polynomials relatively prime? *J. Symb. Comput.* (1998) **267**, 677–689.

[KS97] F. Kako and T. Sasaki. Proposal of "effective floating-point number" for approximate algebraic computation. *preprint*, 10 pages (1997).

[Sas00] T. Sasaki. Mechanism of cancellation errors in multivariate Hensel construction with floating-point numbers. *preprint*.

[SI00] T. Sasaki and D. Inaba. The extended Hensel construction and factorization of multivariate polynomials (Japanese). Research on the theory and applications of computer algebra (Japanese) (Kyoto, 1999). *Surikaisekikenkyusho Kokyuroku* (2000) **1138**, 28–42.

[SK99] T. Sasaki and F. Kako. Solving multivariate algebraic equation by Hensel construction. *Japan J. Indust. Appl. Math.* (1999) **16**, No. 2, 257–285.

[SY98] T. Sasaki and S. Yamaguchi. An Analysis of Cancellation Error in Multivariate Hensel Construction with Floating-point Number Arithmetic. *Proc. ACM Internat. Symp. on Symbolic and Algebraic Computation* (1998), 1–8.