Evaluating Signs of Determinants Using Single-Precision Arithmetic

Jean-Daniel Boissonnat
INRIA Sophia-Antipolis

May 15, 1995

[summary by Brigitte Vallée, Université de Caen]

Abstract

Most decisions in geometric algorithms are based on signs of determinants. For example, deciding if a point belongs to a given half-space or a given ball reduces to evaluating the sign of a determinant. It is therefore crucial to have reliable answers to such tests and to produce robust algorithms. There exist basically two categories of approaches to this objective:

- rounded computations, followed by a proof of the topological correctness of the result;
- exact integer computations that use $n\ell$ bits for the computation of an $n \times n$ determinant with ℓ-bit integers as inputs.

Here, the second approach is followed, the goal being to use as few bits as possible to evaluate signs of determinants. For dimensions $n = 2$ and $n = 3$, the algorithms proposed require respectively ℓ and $\ell + 1$ bits arithmetic, and run in polynomial time in ℓ: they perform in the worst case respectively $O(\ell)$ and $O(\ell^3)$ elementary operations —additions, subtractions, comparisons, and Euclidean divisions— on integers of ℓ or $\ell + 1$ bits. Extensive simulations have shown that the algorithms perform well in practice so that the average-case complexity appears to be much better than the worst-case complexity. This observation can be proved in the two-dimensional case [6]. Under heuristic hypotheses, the proof can be extended to the three-dimensional case [5].

This talk is based on a joint paper of Francis Avnaim, Jean-Daniel Boissonnat, Olivier Devillers, Franco P. Preparata, and Mariette Yvinec [1]. The author of the summary has interpreted some ideas of the original lecture and has proven a conjecture stated there [5, 6].

1. Two-dimensional case.

The aim is to evaluate the sign of a 2×2 nontrivial determinant,

$$D = \det \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \end{pmatrix},$$

with nonzero integer entries of at most ℓ bits. By dividing the first column by x_1 and the second column by y_1, one can write $D = x_1 y_1 D'$ with

$$D' = \det \begin{pmatrix} 1 & 1 \\ x & y \end{pmatrix} = y - x,$$

where $y = y_2/y_1$ and $x = x_2/x_1$. Evaluating the sign of a 2×2 determinant thus reduces to evaluating the sign of the difference between two rationals x and y. We consider the set $\mathcal{J}_\ell$ of rationals in the interval $]0, 1/2]$ whose numerator and denominator have at most ℓ bits.

The main idea is to expand both rationals x and y into continued fractions: the comparison between both expansions (under lexicographic order) suffices to compare the rationals themselves. The outline of the algorithm is thus very simple:

> As long as x and y have matching continued fractions expansions, continue expanding; stop as soon as the expansions differ.

There are two variants of the algorithm, which depend on the kind of continued fraction that is used: The *Standard-Sign* algorithm is based on standard continued fractions that are built with the usual Euclidean division $a = bq + r$ with $0 \leq r < b$ while the *Centered-Sign* algorithm is based on centered continued fractions that are built with the centered Euclidean division $a = bq + r$ with $|r| \leq b/2$. The worst-case of these algorithms arises when x and y are equal and the analysis uses well-known results of Lamé (1845) [4] and Dupré (1846) [3] relative to the standard and centered gcd algorithm respectively. The average number of iterations is quite different from the worst case since it is asymptotically constant (i.e., independent of the number ℓ of bits of the input) [6]. Not too surprisingly similar constants show up in the average-case analysis of lattice reduction algorithms in the two-dimensional case [2].

THEOREM 1. *On rationals x and y of $\mathcal{J}_\ell$, the algorithms perform a number of iterations L at most*

$$\begin{cases} \ell \frac{\log 2}{\log(1+\sqrt{2})}, & \text{for the Centered-Sign algorithm,} \\ \ell \frac{\log 2}{\log \phi}, & \text{for the Standard-Sign algorithm.} \end{cases}$$

(Here, ϕ is the golden ratio equal to $\phi = (1+\sqrt{5})/2$). If the entries x and y are taken in the square $\mathcal{J}_\ell \times \mathcal{J}_\ell$ with a density $F(x,y)$ proportional to $|x-y|^r$ (with $r > -1$), the average number of iterations $E[L]$ of the Centered-Sign algorithm is asymptotic to

$$\beta(r) = \frac{4}{\zeta(4+2r)} \sum_{d=1}^{\infty} \frac{1}{d^{2+r}} \sum_{c=\lceil d\phi \rceil}^{\lfloor d\phi^2 \rfloor} \frac{1}{c^{2+r}}, \qquad \ell \to \infty,$$

and the average number of iterations $E[L]$ of the Standard-Sign algorithm is asymptotic to

$$\alpha(r) = \frac{4}{\zeta(4+2r)} \sum_{d=1}^{\infty} \frac{1}{d^{2+r}} \sum_{d < c \leq 2d} \frac{1}{c^{2+r}}, \qquad \ell \to \infty.$$

In particular, when the density F is uniform, the average numbers of iterations are respectively asymptotic to

$$\beta := \beta(0) = \frac{4}{\zeta(4)} \sum_{d=1}^{\infty} \frac{1}{d^2} \sum_{c=\lceil d\phi \rceil}^{\lfloor d\phi^2 \rfloor} \frac{1}{c^2} = 1.08922\ldots,$$

$$\alpha := \alpha(0) = \frac{4}{\zeta(4)} \sum_{d=1}^{\infty} \frac{1}{d^2} \sum_{d < c \leq 2d} \frac{1}{c^2} = 1.20226\ldots.$$

In Fig. 1, the domains $[L = k]$ relative to the Standard-Sign algorithm are represented alternatively in black (for odd values of k) and white (for even values of k).

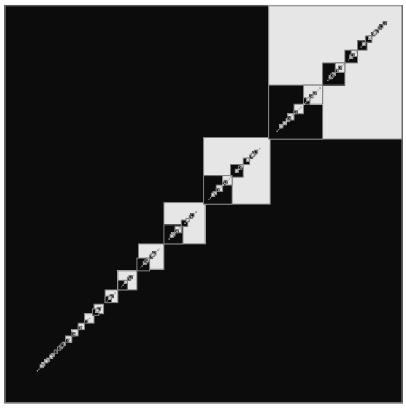

FIGURE 1. The domains $[L = k]$ relative to the Standard-Sign algorithm.

2. Three-dimensional case.

Let D denote a 3×3 determinant with V_1, V_2, V_3 as row vectors. The components x_i, y_i, z_i of vector V_i are assumed to be ℓ-bit integers. The vertical components of the input vectors play a special rôle in the algorithm: all the projections performed are projections parallel to E_z where E_z denotes the last vector of the canonical basis of $\mathbb{R}^3$. Using properties of the determinant, one can assume without loss of generality that the vertical components z_i are positive and in increasing order.

The sign of D depends on which side vector V_3 lies with respect to the plane $\mathcal{P} := \langle V_1, V_2 \rangle$ and this sign is more difficult to evaluate if vector V_3 is very "close" to plane $\mathcal{P}$. We thus define a neighbourhood $\mathcal{V}$ of the plane $\mathcal{P}$ that satisfies the following two properties:

- (a) If V_3 does not belong to $\mathcal{V}$, then it is easy to determine on which side vector V_3 lies with respect to the plane $\mathcal{P}$, the problem reducing to evaluating the sign of a 2×2 determinant;
- (b) If V_3 belongs to $\mathcal{V}$, then there exists a vector W_3 obtained by translating V_3 parallely to $\mathcal{P}$ whose last component z'_3 satisfies $|z'_3| \leq z_3/2$; the algorithm then continues with the new vector system (V_1, V_2, W_3).

The vector V_3 decomposes as $V_3 = \lambda_1 V_1 + \lambda_2 V_2 + \rho E_z$, with rational components $\lambda_1, \lambda_2, \rho$. The numerators and denominators of these rationals may have 2ℓ bits, so that we cannot directly operate with them. The determinant D satisfies

$$D := \det(V_1, V_2, V_3) = \rho \det(V_1, V_2, E_z) = \rho \det(v_1, v_2),$$

where v_i is the projection of V_i on the horizontal plane $z = 0$. If we evaluate the sign of the rational ρ —without explicitly computing it— the problem reduces to evaluating the sign of the determinant $\det(v_1, v_2)$, which is of order 2.

Let $\mathcal{R}$ be the lattice generated by the vectors V_1 and V_2. The fundamental centered parallelogram $\mathcal{F}$ of lattice $\mathcal{R}$ is defined as the set $\mathcal{F} := \{V = \mu_1 V_1 + \mu_2 V_2; |\mu_i| \leq 1/2\}$. The parallelogram $\mathcal{F}$ divides into four sub-parallelograms $\mathcal{F}_i$ that correspond to the four quadrants determined by the four possible signs of (μ_1, μ_2).

Let V be the vector of lattice $\mathcal{R}$ for which $V_3 - V$ is projected inside the fundamental centered

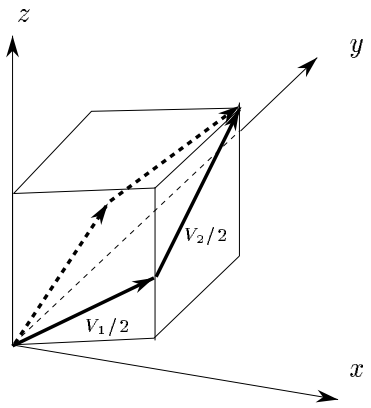

FIGURE 2. The sub-box $\mathcal{B}_1$.

parallelogram $\mathcal{F}$. The integer vector V decomposes as $V := \lfloor\lambda_1\rceil V_1 + \lfloor\lambda_2\rceil V_2$ where $\lfloor\lambda\rceil$ denotes the integer nearest to rational λ. The integer vector $V_3' := V_3 - V$ is $V_3' = \rho_1 V_1 + \rho_2 V_2 + \rho E_z$. Here, the components ρ_i are rationals with absolute value less than $1/2$ and the integer z_3' denotes the vertical component of V_3'. The vector $\rho_1 V_1 + \rho_2 V_2$ is thus a vector of plane $\mathcal{P}$ with rational vertical component $z' := \rho_1 z_1 + \rho_2 z_2$. Since $\rho = z_3' - z'$, the sign of ρ is easy to evaluate provided that we can evaluate the sign of the difference between the integer z_3' and the rational z'. This is the case when the vector V_3' does not belong to the box $\mathcal{B}$ which is defined as follows: The elementary box $\mathcal{B}$ is the union of the four sub-boxes $\mathcal{B}_i$; the sub-box $\mathcal{B}_i$ is a cylinder of direction E_z, of basis $\mathcal{F}_i$, which is delimited by two horizontal planes, whose equations are:

$$\begin{cases} z = 0 \quad \text{and} \quad z = (z_1 + z_2)/2, & \text{for } i = 1; \\ z = -z_1/2 \quad \text{and} \quad z = z_2/2, & \text{for } i = 2; \\ z = -(z_1 + z_2)/2 \quad \text{and} \quad z = 0, & \text{for } i = 3; \\ z = -z_2/2 \quad \text{and} \quad z = z_1/2, & \text{for } i = 4. \end{cases}$$

Thus, each sub-box has an height equal to $(z_1 + z_2)/2$; the sub-box $\mathcal{B}_1$ is represented in Fig 2. The neighborood $\mathcal{V}$ of plane $\mathcal{P}$ is defined as the union of all boxes of the lattice $\mathcal{R}$ obtained by translation of $\mathcal{B}$ by a vector of lattice $\mathcal{R}$.

If the vector V_3 belongs to the neighbourhood $\mathcal{V}$, the vector V_3' belongs to the box $\mathcal{B}$ and different cases are to be considered:

- If the vector V_3' is projected inside $\mathcal{F}_i$ for $i = 2$ or $i = 4$, we let $W_3 := V_3'$; the absolute value of the last component z_3' of vector W_3 is less than $z_2/2$, and the other components of vector W_3 have always at most ℓ bits; the algorithm continues with the system (V_1, V_2, W_3).
- If the vector V_3' is projected inside $\mathcal{F}_i$ for $i = 1$ or $i = 3$, two sub-cases are to be considered: $|\rho_1| \leq |\rho_2|$ and $|\rho_1| \geq |\rho_2|$. The vector W_3 is then defined as follows: W_3 is the vector among the two vectors V_3' or $V_2 - V_3'$ whose last component has the smaller modulus (in the first case). W_3 is the vector among the two vectors V_3' or $V_1 - V_3'$ whose last component has the smaller modulus (in the second case). In both cases, the last component of vector W_3 is in absolute value less than $z_2/2$ and its other components have always at most ℓ bits; the algorithm continues with the system (V_1, V_2, W_3).

We give now the precise description of the algorithm.

Preliminary step. Order the vectors V_1, V_2, V_3 and, if necessary, change some of them into their opposite, in such a way that the vertical components (z_1, z_2, z_3) are positive and sorted in increasing order.

While $z_2 \neq 0$ **do**

1. Compute the vector V of lattice $\mathcal{R}$. Let $V_3' := V_3 - V$.
2. **If** V_3' does not belong to box $\mathcal{B}$
   then evaluate the sign of $\det(v_1, v_2)$ and exit;
   else compute the vector W_3; $V_3 := W_3$. Order the vectors V_1, V_2, V_3 and, if necessary, change them into their opposite so that their vertical components are positive and in increasing order.

At each iteration, either the algorithm evaluates the sign of ρ (if the test in step 2 is positive) and then terminates by the evaluation of a 2×2 determinant, or it continues iteratively on a 3×3 determinant where the largest of the last components has been divided by at least two, the other components remaining unchanged. Thus, the number of iterations of the algorithm is at most equal to 3ℓ.

At each iteration, one computes the nearest integers to rational numbers λ_1 and λ_2,

$$\lambda_1 = \frac{\det \begin{pmatrix} x_3 & y_3 \\ x_2 & y_2 \end{pmatrix}}{\det \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \end{pmatrix}}, \qquad \lambda_2 = \frac{\det \begin{pmatrix} x_1 & y_1 \\ x_3 & y_3 \end{pmatrix}}{\det \begin{pmatrix} x_1 & y_1 \\ x_2 & y_2 \end{pmatrix}}.$$

These rational numbers are quotients of two determinants of order 2 having ℓ-bit integer entries, and cannot be computed directly in single precision. The nearest integers $\lfloor \lambda_i \rceil$ are evaluated, bit by bit by means of a dichotomic process that uses the signs of determinants

$$\det \begin{pmatrix} x_3 - kx_1 & y_3 - ky_1 \\ x_2 & y_2 \end{pmatrix} \quad \text{or} \quad \det \begin{pmatrix} x_3 - kx_2 & y_3 - ky_2 \\ x_1 & y_1 \end{pmatrix}, \qquad |k| = 1, 2, 4, \ldots, 2^\ell,$$

with entries of at most $\ell + 1$ bits. Thus, the computation of vector V uses at most 4ℓ evaluations of signs of 2×2 determinants with entries of at most $\ell + 1$ bits.

THEOREM 2. *Let D be a 3×3 determinant with ℓ-bit integer entries. The determinant sign algorithm above performs at most 3ℓ iterations in the worst-case, each iteration involving the evaluation of at most $4\ell + 9$ signs of determinants 2×2 with $\ell + 1$–bit integer entries. In the worst-case, the algorithm requires $3\ell^2(4\ell + 9)$ elementary steps, each of them involving $O(1)$ additions, comparisons and Euclidean divisions on $\ell + 1$–bit integers.*

Note that extensive experiments show that the average number of iterations is around one. One may give a heuristic explanation to this phenomena [5]. Note also that the algorithm can be generalized to higher dimensions [5].

Bibliography

[1] Avnaim (F.), Boissonnat (J.-D.), Devilliers (O.), F. (Preparata), and Yvinec (M.). – *Evaluating signs of determinants using single-precision arithmetics.* – Research Report n° 2306, Institut National de Recherche en Informatique et en Automatique, 1994.

[2] Daudé (Hervé), Flajolet (Philippe), and Vallée (Brigitte). – An analysis of the Gaussian algorithm for lattice reduction. In Adleman (L.) (editor), *Algorithmic Number Theory Symposium, Lecture Notes in Computer Science*, pp. 144–158. – 1994. Proceedings of *ANTS'94*.

[3] Dupré (A.). – Sur le nombre de divisions à effectuer pour obtenir le plus grand commun diviseur entre deux nombres entiers. *Journal de Mathématiques*, vol. 11, 1846, pp. 41–64.

[4] Lamé (G.). – Note sur la limite du nombre de divisions dans la recherche du plus grand commun diviseur entre deux nombres entiers. *Comptes-Rendus de l'Académie des Sciences*, vol. XIX, 1845, pp. 867–870.
[5] Vallée (B.). – *Algorithme probabiliste pour l'évaluation du signe d'un déterminant $n \times n$ en quasi-simple précision.* – Research report, GREYC, Université de Caen, 1995.
[6] Vallée (B.). – *Evaluation du signe d'un déterminant 2×2: analyse en moyenne.* – Research report, GREYC, Université de Caen, 1995.