Lecture X
DETERMINANTS

Perhaps the most important class of expressions in geometric computation is the determinant. We examine some algorithmic techniques needed for computing determinants in the context of exact geometric computation.

The main issue is to compute some approximation to the determinant which has the correct sign. Clarkson [4] gave a general method for $n \times n$ matrices with L-bit integer entries, assuming limited precision exact integer arithmetic. The method can be very efficient when combined with machine floating point arithmetic for n and L that are moderate. When $n = 2$ or 3, Avnaim et al [1] gave efficient solutions when the inputs are machine representable. In [2], Brönnimann and Yvinec revisit Clarkson's algorithm and generalizes the method of Avnaim et al.

Filters for Determinants. If we use a standard Gaussian elimination algorithm for computing the determinant, then a simple filter is just to compare the absolute value of the computed determinant with the well-known error bounds found in Numerical Analysis books. Pan and Yanquiang [7] pointed out that this filter (which is static) is not very efficacious. They propose instead to use the the distance from the nearest singular instance of a matrix. This requires some extra but not too expensive computation, but leads to much better filtering ratios.

§1. Concepts from Numerical Analysis

In this section, we recall some basic concepts developed in numerical analysis. We assume vectors in $\mathbb{C}^n$ and matrices in $\mathbb{C}^{m\times n}$ for various natural numbers $n, m \geq 1$. For a matrix $A \in \mathbb{C}^{m\times n}$, then $A^* \in \mathbb{C}^{n\times m}$ denotes its Hermitian transpose. In the special case where A is real, A^* is equal to A^T, the transpose. Let $|A|$ be the matrix whose (i, j)th entry is $|a_{ij}|$. Two real matrices A, B of the same dimension can be compared **componentwise**: we write $A \leq B$ if $a_{ij} \leq b_{ij}$ for all i, j. The identity matrix is denoted I (of appropriate dimension determined by context).

References: John [6], Higham [5], Chaitin-Chatelin and Frayssé [3].

Vector Norms. We want some measure of "size" of a vector $x \in \mathbb{C}^n$. This is needed, for instance, in estimating computational errors. This is captured by the notion of a norm. A **norm** on $\mathbb{C}^n$ is a function $N : \mathbb{C}^n \to \mathbb{R}$ such that for all $x, y \in \mathbb{C}^n$,

- $N(x) \geq 0$, with equality iff $x = 0$.
- $N(\alpha x) = |\alpha| N(x)$ for all $\alpha \in \mathbb{C}$.
- $N(x + y) \leq N(x) + N(y)$

The main examples are the p**-norms** for any positive p. This is denoted $N(x) = \|x\|_p$ and defined as

$$\|x\|_p := \sqrt[p]{|x_1|^p + \cdots + |x_n|^p}$$

where $x = (x_1, \ldots, x_n)^T$. The special cases of $p = 1$ and $p = 2$ are noteworthy. In particular, the 2-norm is differentiable and invariant under an orthogonal transformation of the space. We can also generalize this to $p = \infty$ and define

$$\|x\|_\infty := \max\{|x_1|, \ldots, |x_n|\}.$$

Two norms N and N' are **equivalent** if there exists positive constants $c < C$ such that for all $x \in \mathbb{C}^n$,

$$c \cdot N(x) \leq N'(x) \leq C \cdot N(x).$$

It is easy to see that this notion of equivalence is transitive.

Lemma 1 *Any two vector norms are equivalent.*

Proof. By transitivity of equivalence, it suffices to show that any norm N is equivalent to the ∞-norm. Write $x = \sum_{i=1}^n x_i e^i$ where e^i is the ith elementary vector. Then

$$N(x) \leq \sum_{i=1}^n |x_i| N(e^i) \leq \|x\|_\infty \sum_{i=1}^n N(e^i) \leq C\|x\|_\infty$$

where $C = \sum_{i=1}^n N(e^i)$. Now, consider the unit sphere under the ∞-norm, $S = \{x \in \mathbb{C}^n : \|x\|_\infty = 1\}$. Since S is compact, the norm function $N : S \to \mathbb{R}$ achieves its mininum value at some $x^0 \in S$. Let $N(x^0) = c$. If $\|x\|_\infty = b$ then we have

$$N(x) \geq N(b \cdot \frac{x}{b}) = bN(\frac{x}{b}) \geq b \cdot c = c\|x\|_\infty.$$

We have thus specified the requisite constants c, C in the definition of equivalence. **Q.E.D.**

Matrix Norms. We can treat a matrix $A \in \mathbb{C}^{m\times n}$ as a mn-vector and apply the previous vector norms. In particular, the vector 2-norm when applied to A in this way is called the Frobenious norm, and given the special notation $\|A\|_F$. If we view A as a linear operator from $\mathbb{C}^n$ to $\mathbb{C}^m$, then we can use a general method for defining norms on linear operators. To do this, we fix two vector norms N^m and N^n on $\mathbb{C}^m$ and $\mathbb{C}^n$, respectively. We define the **derived norm** N' on $A \in \mathbb{C}^{m\times n}$ as follows:

$$N'(A) := \sup_{x\neq 0, x\in\mathbb{C}^n} \frac{N^m(Ax)}{N^n(x)}. \tag{1}$$

Since $N(\lambda x) = |\lambda| \cdot N(x)$, we can equivalently define N' as

$$N'(A) := \sup_{N^n(x)=1} N^m(Ax). \tag{2}$$

Thus N' measures the maximum "expansion" of a vector on the unit sphere $\{x \in \mathbb{C}^n : N^n(x) = 1\}$ in $\mathbb{C}^n$. We say that N' is **derived** from N^m, N^n. The following properties are easy to show:

1. $N'(A) \geq 0$ with equality iff $A = \mathbf{0}$.
2. $N'(\alpha A) = |\alpha| \cdot N'(A)$ for $\alpha \in \mathbb{C}$.
3. $N'(A + B) \leq N'(A) + N'(B)$.
4. $N'(AB) \leq N'(A)N'(B)$.
5. $N^m(Ax) \leq N'(A)N^n(x)$.

The first 3 properties verify that N' is really a norm. Note that in property 4, up to three derived norms involved! The last property shows a connection between N' and the norms it is derived from. It is customary to simply use 'N' for all three norms, N', N^m and N^n. This is unambiguous as the context will tell us which is the case. With this convention, it is almost mechanical[1] to write down the above 5 properties.

Assume that N^n and N^m are vector p-norms. We want to characterize the derived matrix norm, denoted N_p, for $p = 1, 2, \infty$.

Theorem 2 *Let $A = (a_{ij}) \in \mathbb{C}^{m\times n}$. Assume $B := A^*A$ has eigenvalues $\mu_1, \mu_2, \ldots, \mu_n$.*

$$\begin{array}{llcl} (a) & N_\infty(A) & = & \max_i \sum_j |a_{ij}| \\ (b) & N_1(A) & = & \max_j \sum_i |a_{ij}| \\ (c) & N_2(A) & = & \sqrt{\max_i \mu_i} \end{array}$$

Proof. Let $x = (x_1, \ldots, x_n)^T \in \mathbb{C}^n$ be a non-zero vector.
(a) We have

$$\begin{aligned} \|Ax\|_\infty &= \max_i |\sum_j a_{ij}x_j| \\ &\le \|x\|_\infty \max_i \sum_j |a_{ij}|. \end{aligned}$$

The inequality becomes an equality if we choose $x_j = |a_{i_0 j}|/a_{i_0 j}$ where i_0 is the index that achieves the maximum in $\max_i \sum_j |a_{ij}|$. This proves $N_\infty(A) = \max_i \sum_j |a_{ij}|$. (b) Again, we have

$$\begin{aligned} \|Ax\|_1 &= \sum_i |\sum_j a_{ij}x_j| \\ &\le_{(1)} \sum_i \sum_j |a_{ij}x_j| \\ &\le_{(2)} \sum_j |x_j| \sum_i |a_{ij}x_j| \\ &\le_{(3)} \|x\|_1 \max_j \sum_i |a_{ij}|. \end{aligned}$$

Suppose j_0 is the index that maximizes the expression $\max_j \sum_i |a_{ij}|$, and we choose $x = e^{j_0}$ (the unit vector with 1 in the j_0th position). Then inequalities $\le_{(1)}$ and $\le_{(2)}$ becomes equalities because x is now a unit vector, while inequality $\le_{(2)}$ becomes an equality because of our choice of j_0. This proves $N_1(A) = \max_j \sum_i |a_{ij}|$.
(c) For any μ and x, if $Bx = \mu x$ then $\mu x^* x = x^* Bx = (x^*A^*)(Ax)$. Hence

$$\mu = \frac{\|Ax\|_2^2}{\|x\|_2^2}.$$

This shows that the eigenvalues μ_i of B are real and non-negative. It is well-known that for any Hermitian matrix B (i.e., $B^* = B$), there exists unitary a matrix T (i.e., $T^*T = I$) such that $B = T^*DT$ where D is the diagonal matrix $\mathtt{Diag}(\mu_1, \ldots, \mu_n)$. Without loss of generality, assume μ_1 is the largest eigenvalue. For any x where $\|x\|_2 = 1$, we have

$$\begin{aligned} \|Ax\|_2^2 &= (Ax)^*(Ax) \\ &= x^*Bx \\ &= x^*T^*DTx \end{aligned}$$

[1] We rely on the fact that these properties fit familiar "mathematical patterns" such the triangular inequality.

$$\begin{aligned} &= y^* D y \qquad (y = Tx) \\ &= \sum_i \mu_i |y_i|^2 \\ &\leq \mu_1 \qquad (\|x\|_2 = \|y\|_2 = 1). \end{aligned}$$

The final inequality is an equality if $y = e^1$ (the unit vector with 1 in the first component). This proves $N_2(A) = \sqrt{\max_i \mu_i}$. **Q.E.D.**

We read (a) and (b) as saying: $N_\infty(A)$ is the maximum of the 1-norm of the *rows* of A; $N_1(A)$ is the maximum of the 1-norm of the *columns* of A. To remember whether it is row or column, use this device: the symbol "∞" has a dominant horizontal direction, suggesting rows; likewise, the symbol "1" suggests columns.

It is not easy to compute $N_2(A)$. Hence, the following connection between $N_2(A)$ and the Frobenious norm $\|A\|_F$ of A is useful.

Lemma 3

$$\sqrt{n} N_2(A) \geq \|A\|_F \geq N_2(A).$$

Proof. Note that

$$\|A\|_F^2 = \sum_i \sum_j |a_{ij}|^2 = \mathtt{Tr}(A^* A).$$

Writing $B = A^* A = T^* D T$ as in the above proof, we see

$$\mathtt{Tr}(B) = \sum_i \mu_i.$$

Therefore

$$n \max_i \mu_i \geq \mathtt{Tr}(B) \geq \max_i \mu_i.$$

The result follows by taking square-roots. **Q.E.D.**

These bounds on $N_2(A)$ can be computed in $O(n^2)$ arithmetic operations.

§2. Elements of Error Analysis

Let $\mathbf{u}$ be the unit roundoff for the floating point number system under discussion (see the chapter on floating point numbers). In the floating point system $F(\beta, t)$, $\mathbf{u} = \beta^{1-t}/2$. Recall that the standard error model says that the approximate value $\widetilde{x}$ of any floating point arithmetic operation is related to the exact value x by the relation

$$\widetilde{x} = x(1 + \delta), \qquad |\delta| \leq \mathbf{u}.$$

We develop some basic tools for the error analysis within this model. The following is a slight extension of Higham [5, p. 69].

In the following, we assume that n is small enough so that

$$n\mathbf{u} < 1. \tag{3}$$

For any $n \geq 1$ and real ε, define

$$\gamma_n(\varepsilon) := \frac{n\varepsilon}{1 - n\varepsilon}. \tag{4}$$

When $\varepsilon = \mathbf{u}$, we simply write

$$\gamma_n := \gamma_n(\mathbf{u}).$$

Note that

$$1 + \gamma_n(\varepsilon) = \frac{1}{1 - n\varepsilon} \tag{5}$$

Lemma 4
(i) $\gamma_n(\varepsilon)$ is a strictly increasing function for $\varepsilon < 1/n$.
(ii) If $1 > 2n\mathbf{u}$ then $|\gamma_n(\varepsilon)| \leq \gamma_n$ iff $\varepsilon \in [\frac{-\mathbf{u}}{1-2n\mathbf{u}}, \mathbf{u}]$.
(iii) Assume $1 > 2n\mathbf{u}$. For any real θ, we have $|\theta| \leq \gamma_n$ iff $\theta = \gamma_n(\varepsilon)$ for some $\varepsilon \in [\frac{-\mathbf{u}}{1-2n\mathbf{u}}, \mathbf{u}]$.

Proof. (i) If $\varepsilon < \varepsilon' < 1/n$, then we verify that $1 + \gamma_n(\varepsilon) = \frac{1}{1-n\varepsilon} < \frac{1}{1-n\varepsilon'} = 1 + \gamma_n(\varepsilon')$. Note that $\varepsilon, \varepsilon'$ can be negative.
(ii) Since $1 > 2n\mathbf{u}$, we have $\frac{-\mathbf{u}}{1-2n\mathbf{u}} < 0 < \mathbf{u}$. We check that $\gamma_n(\frac{-\mathbf{u}}{1-2n\mathbf{u}}) = -\gamma_n(\mathbf{u}) = -\gamma_n$. By part (i), it follows that for all $\varepsilon \in [\frac{-\mathbf{u}}{1-2n\mathbf{u}}, \mathbf{u}]$, we have $-\gamma_n \leq \gamma_n(\varepsilon) \leq \gamma_n$.
(iii) This is a consequence of (ii). **Q.E.D.**

Lemma 5 *Let $|\delta_i| \leq \mathbf{u}$, $\rho_i = \pm 1$ and $1 > 2n\mathbf{u}$. Then*

$$\prod_{i=1}^{n} (1 + \delta_i)^{\rho_i} = 1 + \gamma_n(\varepsilon)$$

where for some ε such that $|\gamma_n(\varepsilon)| \leq \mathbf{u}$.

Proof. Suppose $n = 1$. Then

$$\begin{aligned}
(1+\delta_1)^{\rho_1} &= \begin{cases} 1 + \delta_1 & \text{if } \rho_1 = 1, \\ 1 + \frac{-\delta_1}{1+\delta_1} & \text{if } \rho_1 = -1 \end{cases} \\
&= \begin{cases} 1 + \gamma_1\left(\frac{\delta_1}{1+\delta_1}\right) & \text{if } \rho_1 = 1, \\ 1 + \gamma_1(-\delta_1) & \text{if } \rho_1 = -1 \end{cases} \\
&= 1 + \gamma_1(\varepsilon)
\end{aligned}$$

where $\varepsilon \in [\frac{-\mathbf{u}}{1-2\mathbf{u}}, \mathbf{u}]$. Thus $|\gamma_1(\varepsilon)| \leq \mathbf{u}$, as required. Let $n \geq 2$. By induction,

$$\sum_{i=1}^{n-1} (1 + \delta_i)^{\rho_i} = 1 + \gamma_{n-1}(\varepsilon)$$

and

$$(1 + \delta_n)^{\rho_n} = 1 + \gamma_1(\varepsilon')$$

where $|\gamma_{n-1}(\varepsilon)| \leq \mathbf{u}$ and $|\gamma_1(\varepsilon')| \leq \mathbf{u}$. It suffices to show that

$$1 + \gamma_n(\varepsilon'') = (1 + \gamma_{n-1}(\varepsilon))(1 + \gamma_1(\varepsilon')) \tag{6}$$

for some ε'' satisfying $|\gamma_n(\varepsilon'')| \le \mathbf{u}$. Alternatively, we must show that ε'' lies in the range $[-\mathbf{u}/(1-2n\mathbf{u}), \mathbf{u}]$, given that $\varepsilon \in [-\mathbf{u}/(1-2(n-1)\mathbf{u}), \mathbf{u}]$ and $\varepsilon' \in [-\mathbf{u}/(1-2\mathbf{u}), \mathbf{u}]$. It follows from (5) that (6) amounts to showing

$$\begin{aligned} \frac{1}{1-n\varepsilon''} &= \frac{1}{1-(n-1)\varepsilon} \cdot \frac{1}{1-\varepsilon'} \\ 1-n\varepsilon'' &= (1-(n-1)\varepsilon)(1-\varepsilon') \\ \varepsilon'' &= \frac{n-1}{n}\varepsilon(1-\varepsilon') + \frac{1}{n}\varepsilon'. \end{aligned}$$

Thus ε'' is a convex combination of $\varepsilon(1-\varepsilon')$ and ε'. There are three cases to consider:
Case A. If $\varepsilon' \ge 0$, then both $\varepsilon(1-\varepsilon')$ and ε' lie in the range $[-\mathbf{u}/(1-2n\mathbf{u}), \mathbf{u}]$. So its convex combination ε'' lies in the range $[-\mathbf{u}/(1-2n\mathbf{u}), \mathbf{u}]$.
Case B. If $\varepsilon' < 0$ and $\varepsilon \ge 0$ then

$$\varepsilon'' \le \frac{n-1}{n}\varepsilon(1-\varepsilon') \le \frac{n-1}{n}\mathbf{u}\left(1 - \frac{\mathbf{u}}{1-2\mathbf{u}}\right) \le \mathbf{u}$$

and

$$\varepsilon'' \ge \frac{1}{n}\varepsilon' \ge \frac{-\mathbf{u}}{n(1-2\mathbf{u})} \ge \frac{-\mathbf{u}}{1-2n\mathbf{u}}.$$

Case C. Finally, assume $\varepsilon' < 0$ and $\varepsilon < 0$. Then $\varepsilon'' < 0$ and we need to show that $\varepsilon'' \ge -\mathbf{u}/(1-2n\mathbf{u})$. Since ε'' is a convex combination of $\varepsilon(1-\varepsilon')$ and ε', it suffices to show that both $\varepsilon(1-\varepsilon')$ and ε' are at least $-\mathbf{u}/(1-2n\mathbf{u})$. Clearly, $\varepsilon' \ge -\mathbf{u}/(1-2\mathbf{u}) \ge -\mathbf{u}/(1-2n\mathbf{u})$. Also,

$$\begin{aligned} \varepsilon(1-\varepsilon') &\ge \varepsilon(1+\mathbf{u}) \\ &\ge \frac{-\mathbf{u}(1+\mathbf{u})}{1-2(n-1)\mathbf{u}} \\ &\ge \frac{-\mathbf{u}}{1-2n\mathbf{u}}, \end{aligned}$$

where the last inequality may be verified directly. **Q.E.D.**

On Evaluation of Expressions. We consider expressions E in the basic arithmetic operations, generalized to allow summation ($\sum_{i=1}^{n}$) or multi-product ($\prod_{i=1}^{n}$). E.g.,

$$E = 3x_2 + x_3 - (x_1x_2x_3) + \sum_{i=1}^{5} x_i.$$

An expression gives a mathematical value, but we also are interested in its evaluation on a computer. In this case, the order of evaluating the expression is important. This order is unique for a fully parenthesized expression. But in general, for any E, we define the **natural evaluation** of E to be the direct left-to-right of expressions, respecting any parentheses when present. Each summation or multiproduct notation is linearized in the obvious way, by increasing indices and evaluated left-to-right. E.g., $\sum_{i=1}^{n} x_n$ is simply a shorthand for $x_1 + x_2 + \cdots + x_n$, whose natural evaluation order is

$$((\cdots((x_1 + x_2) + x_3) + \cdots + x_{n-1}) + x_n)$$

Suppose E' is any **equivalent** expression, *i.e.*, obtained from E by applying any valid laws of arithmetic (associativity, commutativity, distributivity, etc). Because of finite precision arithmetic, the evaluation of E' and E need not produce the same result. The simplest example is

$$E = x + y + z, \qquad E' = x + z + y. \tag{7}$$

Another example is the following, using the law of distributivity:

$$E = (x+y)(v+w), \qquad E' = xv + xw + yv + yw. \tag{8}$$

In numerical analysis, we tend not to exploit distributivity because they can render E' very different from E. Comparing the errors in the natural evaluations E and E' can be complicated. One hint of this is the fact that the number of operations in E and E' can be exponentially different:

$$E = \prod(1+x_i), \qquad E' = 1 + \sum_i x_i + \sum_{i<j} x_i x_j + \cdots \prod_i x_i.$$

Therefore, we focus on one class of mathematically equivalent expressions, called **re-ordering equivalent**. Two expressions are re-ordering equivalent if they can be obtained from each other by applying the laws of associativity and commutativity only. If E is re-ordering equivalent to E', then their natural evaluation incurs the same number of arithetic operations.

An approximate value $\widetilde{E}$ for E is called a **natural value** (resp. **re-ordering value**) if is obtained by the natural evaluation (resp. re-ordering evaluation) of E.

Lemma 6 *Let $E = \sum_{i=1}^n x_i y_i$.*
(i) The natural value $\widetilde{E}$ of E satisfies $\widetilde{E} = \sum_{i=1}^n x_i y_i (1+\theta_i)$ where $|\theta_1| \le \gamma_n$ and for $i = 2, \ldots, n$, $|\theta_i| \le \gamma_{n-i+2}$.
(ii) If E' is any re-ordering equivalent expression, its natural value $\widetilde{E'}$ satisfies $\widetilde{E'} = \sum_{i=1}^n x_i y_i (1+\theta_i)$ where each $|\theta_i| \le \gamma_n$.

Proof. Let $E_j = \sum_{i=1}^j x_i y_i$ for $j = 1, \ldots, n$.
(i) If $n = 1$, $\widetilde{E} = x_1 y_1 (1+\theta_1)$, $|\theta_1| \le \gamma_1$. This has the required form. For $n \ge 2$, we inductively have

$$E^{\rceil}_{n-1} = \sum_{i=1}^{n-1} x_i y_i (1+\theta'_i)$$

where $|\theta'_1| \le \gamma_{n-1}$ and for $i = 2, \ldots, n-1$, $|\theta'_i| \le \gamma_{n-i+1}$. From

$$\begin{aligned}
\widetilde{E_n} &= (E^{\rceil}_{n-1} + x_n y_n (1+\theta'_n))(1+\theta) \\
&= \sum_{i=1}^{n} x_i y_i (1+\theta'_i)(1+\theta) \\
&= \sum_{i=1}^{n} x_i y_i (1+\theta_i)
\end{aligned}$$

where $|\theta_i|$ satisfies the lemma.
(ii) This is immediate from part (i). **Q.E.D.**

Lemma 7 *Let $E = (z - \sum_{i=1}^{n-1} x_i y_i)/y_n$. If E' is any re-ordering equivalent expression, its natural value $\widetilde{E'}$ satisfies*

$$\widetilde{E'}(1+\theta_n) = \frac{z - \sum_{i=1}^{n-1} x_i y_i (1+\theta_i)}{y_n}$$

where each $|\theta_i| \le \gamma_n$.

Proof. **Q.E.D.**

§3. Conditioning

Let $f : X \to Y$ be a continuous function on the normed spaces X, Y. The **condition number** $\kappa_f(x)$ of an **input instance** $x \in X$ for problem f measures the sensitivity of $f(x)$ to perturbations in x. In particular, "ill or well conditioning" refers to the relative largeness or smallness of $\kappa_f(x)$. We can work with an absolute or a relative notion of conditioning. We follow [8] in the following definitions.

There is an absolute and a relative concept of conditioning. The relative notion is more important, but since the absolute notion is simpler, we also will present it. The **absolute conditioning function** of f is defined by

$$\kappa_f^A(x) := \lim_{\delta \to 0} \sup_{\|\delta x\| \le \delta} \frac{\|\delta f\|}{\|\delta x\|}. \tag{9}$$

Here, we write δf for $f(x + \delta x) - f(x)$. If f is differentiable, write $J(x) = J_f(x)$ for the Jacobian of f at x. The (i,j)th entry of $J(x)$ is $\partial f_i / \partial x_j(x)$. Then the quantities δx and δf becomes (respectively) the infinitesimals dx and $J(x)dx$. Thus

$$\kappa_f^A(x) = \|J_f(x)\| \tag{10}$$

where $\|J_f(x)\|$ is the derived norm for the norms on X and Y.

The **relative conditioning function** of f is defined as

$$\kappa_f^R(x) := \lim_{\delta \to 0} \sup_{\|\delta x\| \le \delta} \left(\frac{\|\delta f\|}{\|f(x)\|} \Big/ \frac{\|\delta x\|}{\|x\|}. \right) \tag{11}$$

Again when f is differentiable, we get

$$\kappa_f^R(x) = \|J_f(x)\| \frac{\|x\|}{\|f(x)\|}. \tag{12}$$

In the following, we simply write $\kappa_f(x)$ to refer to the relative conditioning function.

Condition Number of a Matrix. Let $f : \mathtt{C}^n \to \mathtt{C}^m$ where

$$f(x) = Ax$$

where A is a $m \times n$ matrix. Applying the definitions directly, we obtain

$$\begin{aligned}
\kappa_A(x) &:= \lim_{\delta \to 0} \sup_{\|\delta x\| \le \delta} \left(\frac{\|\delta f\|}{\|f(x)\|} \Big/ \frac{\|\delta x\|}{\|x\|} \right) \\
&= \lim_{\delta \to 0} \sup_{\|\delta x\| \le \delta} \left(\frac{\|A(x + \delta x) - Ax\|}{\|Ax\|} \Big/ \frac{\|\delta x\|}{\|x\|} \right) \\
&= \lim_{\delta \to 0} \sup_{\|\delta x\| \le \delta} \left(\frac{\|A\delta x\| \cdot \|x\|}{\|Ax\| \cdot \|\delta x\|} \right) \\
&= \frac{\|A\| \cdot \|x\|}{\|Ax\|}
\end{aligned}$$

If A is square, and A^{-1} exists, then we have $\|x\| = \|A^{-1}Ax\| \le \|A^{-1}\| \cdot \|Ax\|$, Plugging for $\|x\|$ into the above condition number, we obtain

$$\kappa_A(x) \le \|A\| \cdot \|A^{-1}\|.$$

The nice thing about the righthand side is that it is independent of x. This quantity is so useful that it deserves its own name and notation:

$$\kappa(A) := \|A\| \cdot \|A^{-1}\|$$

is the **condition number** of the square matrix A.

This number $\kappa(A)$ is also when we want to solve the linear equation $Ax = b$ where b is given and we need to compute x. This is just the original problem with A^{-1} in place of A, but because of the way we define condition numbers of a matrix, the condition number of A can serve to bound the conditioning of this problem as well.

Condition Number of matrix. This is defined to be

$$\kappa(A) := \|A^{-1}\| \|A\|.$$

Normwise vs. componentwise norm. $|A| = [|a_i j|]$

Distance to a Singularity. The numerical stability of a numerical problem is influenced by its distance to the nearest singularity. For linear systems, this distance can be defined as follows: for a nonsingular matrix A and any matrix norm N, the closest distance from A to a singular matrix is defined as

$$\delta(A) := \inf_S \{N(S - A)\}$$

where S range over all singular matrices. We next characterize this distance. This result, attributed to Turing, was known to Banach in the 1920s. Its generalization to arbitrary norms is from Gastinel (1966). First, we note a simple fact:

Lemma 8 *For any square matrix A and any derived norm N, if $N(A) < 1$ then $I + A$ is invertible.*

Proof. This is equivalent to showing that $(I + A)x = \mathbf{0}$ has the unique solution $x = \mathbf{0}$. If $(I + A)x = \mathbf{0}$ then $x = -Ax$ and hence $N(x) = N(Ax) \le N(A)N(x)$. We claim that $N(x) = 0$ since otherwise $N(A) < 1$ implies $N(x) \le N(A)N(x) < N(x)$. Thus $x = \mathbf{0}$. **Q.E.D.**

Theorem 9 (Turing-Gastinel) *For non-singular A and derived norm N,*

$$\delta(A) = \inf_S \{N(A - S)\} = \frac{1}{N(A^{-1})}.$$

Proof. We first prove that $\delta(A) \ge \frac{1}{N(A^{-1})}$. Let S be any matrix such that $N(S - A) < \frac{1}{N(A^{-1})}$. It suffices to that S is non-singular. Writing Δ for $S - A$, we have $N(A^{-1}\Delta) \le N(A^{-1})N(\Delta) < 1$. Thus $(I + A^{-1}\Delta)$ is invertible and we see that

$$A(I + A^{-1}\Delta) = A + \Delta = S$$

As a product to two non-singular matrices, S is non-singular.

To show that $\delta_T(A) \le \frac{1}{N(A^{-1})N(A)}$, we construct a singular matrix S such that $N(S - A) = \frac{1}{N(A^{-1})}$. We pick two vectors x, y such that

$$y^* A^{-1} x = N(A^{-1}), \qquad N(xy^*) = 1.$$

We check that

$$\left(A - \frac{xy^*}{N(A^{-1})}\right) A^{-1}x = 0.$$

This proves that $S = A - \frac{xy^*}{N(A^{-1})}$ is singular. Moreover, $N(A-S) = N(xy^*)/N(A^{-1}) = 1/N(A^{-1})$. **Q.E.D.**

Component-wise Distance. For a non-singular square matrix A, define (see [3])

$$\delta_T(A) := \inf_S \left\{ \frac{N(S-A)}{N(A)} \right\} \tag{13}$$

where S ranges all singular matrices. Thus $\delta_T(A)$ is the relative distance from A to the nearest singular matrix S. The subscript refers to Turing. Rewrite the definition of $\delta_A(T)$ in (13) as

$$\delta_T(A) = \inf_S \{\alpha : N(S-A) = \alpha N(A)\}$$

where S ranges over all singular matrices. We can further rewrite it as

$$\delta_T(A) = \inf_{S,\alpha} \{\alpha : N(S-A) \le \alpha N(A)\}$$

The componentwise analogue of this definition is therefore

$$\delta_C(A) := \inf_{S,\alpha} \{\alpha : |S-A| \le \alpha |A|\}.$$

A characterization of $\delta_C(A)$ is given by Rohn (1990) (see [3, p.66]).

§4. Convex Hulls and Partial Compilation of Expressions

Sometimes, part of the arguments of an expression are fixed while others are vary. We can exploit this by constructing "partially compiled expressions". To see an example of this phenomenon, we look at the Beneath-Beyond Algorithm for convex hulls.

See [Page 147ff in Edelsbrunner].

Let S be a set of points in d dimensions. FOR SIMPLICITY, ASSUME GENERAL POSITION. We sort the points by their first coordinate, and add them in this order. The method (called Beneath-Beyond) begins with a $d+1$ simplex. At each step, we add a new point p to the current hull H:

Exercises

Exercise 4.1: Prove the 5 properties stated for the norm N' derived from N^m, N^n. ♢

Exercise 4.2: Let A be a square matrix and N a matrix norm.
(i) Show that $N(A) \ge |\mu|$ for any eigenvalue of A.
(ii) Let us show that the above inequality cannot be improved to an equality. Let

$$A = \begin{vmatrix} 1 & 1 \\ 0 & 1 \end{vmatrix}.$$

Then the eigenvalues of A satisfy $|\mu_1| = |\mu_2| = 1$. Prove that for any norm N, we have $N(A) > 1$. Hint: Write $A = I + E$ where I is the identity matrix. What is A^k for any $k \in \mathtt{N}$? Show that if $N(A) \le 1$ then $N(E) = 0$, contradiction.
(iii) (Fritz John) Show that for all square matrices A and all $\varepsilon > 0$, there exists a norm N such that $N(A) \le \varepsilon + \max_\mu |\mu|$ where μ ranges over the eigenvalues of A. ♢

Exercise 4.3: (a) Give specific numerical values for the expressions E and E' (7) so that their natural evaluations result is different answers. Assume IEEE arithmetic.
(b) Do the same for (8). ♢

Exercise 4.4: Improve lemma 5 in case $\rho_i = 1$ for all i. ♢

Exercise 4.5: (Trefethan) Determine the absolute and relative conditioning function of the following functions.
(i) $f(x) = x/2$ for $x \in \mathtt{C}$.
(ii) Suppose $f(x) = \sqrt{x}$ for $x > 0$.
(ii) Suppose $f(x) = x - y$ for $(x, y) \in \mathtt{C}^2$. Use the ∞-norm on $\mathtt{C}^2$. ♢

End Exercises

1 OLD

REFERENCE: Certified Computation of the Sign of a Matrix Determinant, Dario A. Bini, Victor Y. Pan and Yanqiang Yu. 1998, abstract.

An approach to computing the sign of a determinant D of an $n \times n$ matrix M is to numerically compute an approximate value A for D, estimate the roundoff error E, and to certify the computed sign of A in case $|A| > E$. Unfortunately, $|D|$ varies from 0 to $\|M\|^n$, and the the known techniques for error analysis gives error bounds that are large even when $|D|$ is small. A predecessor to this paper by Pan, Yu and Stewart uses this approach. The present paper avoids estimating the roundoff error E, but instead solve the simpler task of estimating the mininum distance $1/\|M^{-1}\|$ from M to a singular matrix. The method uses $O(n^3)$ single precision flops. Based on this estimate, in case the sign is not certified, a decision can be made for increasing the precision of the numerical computation or to fall back to symbolic guaranteed methods. Thus, this algorithm can be used as a kind of filter.

Other work: Clarkson proposed a general algorithm based on Gram-Schmidt orthogonalization. The certification method is rather complicated to program. For $n \le 4$, Avnaim et al proposed an efficient "single-precision" method. Modular methods have been proposed by Brönnimann et al.

2 Roundoff Errors in Gaussian Elimination

Let A be a $n \times n$ matrix, and A_k be its kth column and $a_{i,j}$ its (i,j)th entry. The following matrix operator norms are used: the 2-norm $\|A\|_2$, the row norm $\|A\|_\infty = \max_i \sum_j |a_{i,j}|$, the column norm $\|A\|_1 =$

$\max_j \sum_i |a_{i,j}|$. Also, let $|A|$ denote the matrix with (i,j)th entry $|a_{i,j}|$. A triangular matrix is **unit** (resp., **proper**) if its diagonal entries are all 1 (resp., 0).

Theorem 10 *(Higham, Theorem 9.3, page. 175): Let*

$$A = PA'Q, \quad \widetilde{A} = A' + E = \widetilde{L}\widetilde{U}$$

where the matrices are all n square, P and Q are permutation matrices, $\widetilde{L}$ *and* $\widetilde{U}$ *are unit lower triangular and U upper triangular. If* $\widetilde{L}$ *and* $\widetilde{U}$ *are computed numerically from A by Gaussian elimination with complete pivoting, then*

$$|E| \le \gamma_n |\widetilde{L}| \cdot |\widetilde{U}|$$

where $\gamma_n = n\varepsilon/(1-n\varepsilon)$ *and* ε *is the unit roundoff.*

Since $\det(A) = \det(P)\det(A')\det(Q)$ and $\det(P) = \pm 1$ and $\det(Q) = \pm 1$ are known, it is enough to determine the sign of $\det(A')$.

Corollary 11 *For any operator norm,*

$$\|E\| \le e = \|(\|\widetilde{L}\| \cdot \|\widetilde{U}\|)\|\gamma_n$$

Computing the columns and row norms uses $O(n^2)$ arithmetic operations.

References

[1] F. Avnaim, J.-D. Boissonnat, O. Devillers, F. Preparata, and M. Yvinec. Evaluating signs of determinants using single-precision arithmetic. *Algorithmica*, 17(2):111–132, 1997.

[2] H. Brönnimann and M. Yvinec. Efficient exact evaluation of signs of determinants. *Algorithmica*, 27:21–56, 2000.

[3] F. Chaitin-Chatelin and V. Frayssé. *Lectures on Finite Precision Computations.* Society for Industrial and Applied Mathematics, Philadelphia, 1996.

[4] K. L. Clarkson. Safe and effective determinant evaluation. *IEEE Foundations of Computer Science*, 33:387–395, 1992.

[5] N. J. Higham. *Accuracy and stability of numerical algorithms.* Society for Industrial and Applied Mathematics, Philadelphia, 1996.

[6] F. John. *Lectures on Advanced Numerical Analysis.* Gordon and Breach, 150, Fifth Avenue, New York, NY 10011, 1967.

[7] V. Y. Pan and Y. Yu. Certified computation of the sign of a matrix determinant. *Proc. 10th ACM-SIAM Symp. on Discrete Algorithms (SODA99)*, pages 715–724, 1999. Accepted for Algorithmica.

[8] L. N. Trefethen and I. David Bau. *Numerical Linear Algebra.* Society for Industrial and Applied Mathematics, Philadelphia, 1997.