ADVANCES IN APPLIED MATHEMATICS 1, 1–6 (1980)

The Ellipsoid Method in Linear Programming

BURTON RANDOL

Department of Mathematics, CUNY Graduate Center, 33 West 42nd Street, New York, New York 10036

Recently, Khachian [3] presented a short algorithm for determining the consistency, i.e., solvability, of a finite set of linear inequalities in R^n (cf. also [5]). More precisely, if

$$\begin{aligned} a_{11}x_1 + \dots + a_{1n}x_n &< b_1 \\ a_{m1}x_1 + \dots + a_{mn}x_n &< b_m \\ x_1 > 0, \dots, x_n &> 0 \end{aligned} \qquad (*)$$

is a system of inequalities in R^n for which the a_{ij}'s and b_i's are integers, the algorithm will determine whether or not $(*)$ has a solution, and locate a solution if there is one. Furthermore, it does this in such a way that the required accuracy, memory requirements, and running time are bounded by fixed polynomials in n and L, where L is the complexity of the system as measured by the number of bits needed to encode $(*)$.

Remarks. 1. The requirement that the solution set lie in the positive orthant is a standard one in linear programming, and a problem without this constraint can very easily be replaced by an equivalent one satisfying this requirement (cf. [1, p. 86]).

2. As is shown in [2], the consistency of an integer system in which the inequalities are weak, i.e., of the form "$\leq$" can be made equivalent to that of a system of strict integer inequalities. Moreover, a system of weak integer inequalities can in practice be approximated by a system of rational, and hence integer strict, inequalities. Finally, if a problem involves an auxiliary linear form which is to be maximized subject to a system of weak inequalities, it can be changed into a system without a maximization constraint by adjoining the dual problem, together with the requirement that the forms to be extremalized in the respective problems be equal (cf. [1, p. 129]).

The papers of Khachian and Shor are in the style of brief research announcements and do not contain proofs of several assertions required to

0196-8858/80/010001-06$05.00/0

establish the algorithm. Some proofs have recently been provided in [2]. All of these papers focus mainly on algebraic descriptions of the algorithm, and in the case of [3], on the necessary detailed accuracy, memory, and running time estimates. It is the purpose of this paper to sketch briefly the extremely simple geometric idea involved in the algorithm, and then describe a possible improvement suggested by A. T. Vasquez.

Basically, the technique is very evocative of one of the standard approaches to the Heine–Borel theorem—repetitive bisection. The new idea is that since the bisecting hyperplanes which arise can be tilted in random ways, it is extremely advantageous to make the repeated bisections involve a figure which is invariant under affine transformations, e.g., a half-ellipsoid. In particular, because of this affine invariance, a certain circumscribing lemma has the same geometric form at all stages of the process.

In more detail, after setting $A_{ij} = |a_{ij}| + 1$, $B_i = |b_i| + 1$, and $M =$ (the product of all the A_{ij}'s and B_i's), one begins with the following simple observations:

A. If (∗) is consistent, then the coordinate entries of any vertex are rational numbers whose numerators and denominators are less than M. Thus, if (∗) is consistent, it always has a solution within the open ball B_0 of radius $\sqrt{n}\, M$ about the origin.

B. If (∗) is consistent, the volume of the portion of the solution set which lies in B_0 is greater than $V_0 = (n!)^{-1} M^{-(n+1)}$. (As is remarked in [2], the quantity V_0 is itself of the form $2^{-P(n, L)}$, where $P(n, L)$ is a linear polynomial in n and L.)

A and B are proved in [2], and are also familiar from the geometry of numbers. To prove A, one observes that any coordinate entry of a vertex of the solution polyhedron must be a solution of an n by n integer linear system with entries of the form 0, 1, a_{ij}, and b_i, and hence, is expressible as the quotient of two determinants with such entries. Each determinant can be trivially bounded by replacing it by the product of the sums of the absolute values of the rows, which in turn is dominated by M, which proves A. To prove B, one observes that the solution polyhedron must have at least one vertex, since it is contained in the positive orthant. Such a vertex must, of course, lie in B_0. The solution polyhedron may or may not have $n + 1$ vertices in general position, but if it does not, the problem can be augmented by the additional constraints $x_1 < M, \ldots, x_n < M$ without disrupting the presence of a vertex, and hence, a solution within B_0. These additional constraints cause the solution polyhedron to have compact closure, and hence, certainly produce $n + 1$ vertices in general position in $\bar{B}_0$. The volume of the simplex spanned by these vertices is

$$(n!)^{-1}\left|\det\begin{pmatrix} 1 & \cdots & 1 \\ v_1 & \cdots & v_{n+1} \end{pmatrix}\right|,$$

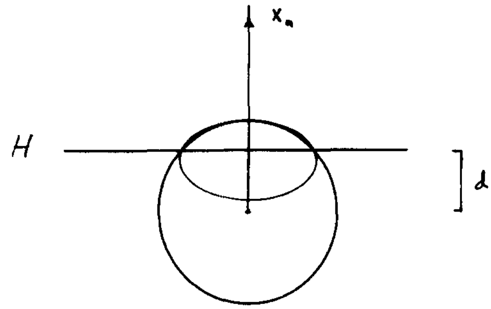

FIGURE 1

where $v_1, \ldots, v_{n+1}$ are column vectors representing the vertices. Since the determinant is a rational number whose numerator is not zero, and since the denominators of the entries of any fixed vertex are all the same and less than M, the result follows immediately, if we combine the formal expansion of the determinant into a single fraction.

The next observation is of a purely geometric character. If E_1 is an open ellipsoid in R^n, and E_1^* denotes the intersection of E_1 with a closed half-space not containing more than half of E_1 and having boundary hyperplane H, then there exists a second open ellipsoid E_2, such that E_2 contains E_1^*, and $(\mathrm{vol}(E_2))/(\mathrm{vol}(E_1)) < e^{-1/2n}$. Since the truth of this assertion is an affine invariant, it suffices to establish it in the case in which E_1 is the unit ball centered at the origin, and H is given by $x_n = d$, with $0 \leq d < 1$ (Fig. 1).

If we require the circumscribing ellipsoid to be rotationally invariant about the x_n axis, it follows quickly from the calculus that the center of the minimal circumscribing ellipsoid is located at distance $(nd + 1)/(n + 1)$ from the origin, and its short semi-axis is $(n/(n + 1))(1 - d)$, while the long semi-axes are identical, and of length $(n^2/(n^2 - 1))^{1/2}(1 - d^2)^{1/2}$. It follows that the ratio of the volume to that of the unit ball in the worst possible case when $d = 0$ is $(n/(n + 1))(n^2/(n^2 - 1))^{(n-1)/2} = r_n$. This quantity is less than 1 if and only if its logarithm is less than zero. After a little manipulation, the statement about logarithms becomes $\log(1 + n^{-1}) + \frac{1}{2}(n - 1)\log(1 - n^{-2}) > 0$, which follows from the power series expansions for $\log(1 + x)$ and $\log(1 - x^2)$. Although $r_n \to 1$ as $n \to \infty$, it follows immediately from the power series expansions that r_n is always less than $e^{-1/2n}$. Finally, if E_1 is specified by giving both its center p_1, and the symmetric matrix E_1' of the quadratic form inverse to the form defining the ellipsoid obtained by parallel translating E_1 to the origin, and if one is given the hyperplane H, then the determination of the center of E_2 and the matrix E_2' is not very time-consuming. Indeed, suppose that T_1 is a linear transformation which takes the unit ball into E_1. (To save notation, we assume that T_1 is symmetric.) Now the hyperplane H will in practice be given by one of the inequalities in (*). If $a = (a_{i1}, \ldots, a_{in})$ is the vector corresponding to this inequality, it is a simple exercise in affine geometry to show that the image under T_1 of the point of tangency of the circumscribing ellipsoid in the ball model is $s\|T_1 a\|^{-1} T_1^2 a$, where $s = \operatorname{sign}(b_i - (a, p_1))$, and hence, the center of E_2 is obtained by displacing the center of E_1 by $s(nd + 1)(n + 1)^{-1}\|T_1 a\|^{-1} T_1^2 a = s(nd + 1)(n + 1)^{-1}(E_1' a, a)^{-1/2} E_1' a$. (The algorithm as stated in [2] has the wrong sign for the displacement vector.)

To find E_2' we proceed as follows. Suppose θ is a unit vector. Denote by $E_\theta^*(d)$ the ellipsoid obtained by translating the circumscribing ellipsoid of Fig. 1 to the origin and then subjecting it to an orthogonal transformation which takes the north pole into θ. It is then evident that $E_\theta^*(d)$ is the image of the unit ball under a transformation of the form $c_1 I + c_2 P_\theta$, where P_θ is the projection along θ, $c_1 = (n^2/(n^2 - 1))^{1/2}(1 - d^2)^{1/2}$, and $c_2 = (n/(n + 1))(1 - d) - c_1$. A transformation taking the unit ball into E_2 will then be given by $T_2 = T_1(c_1 I + c_2 P_\theta)$, which is generally not symmetric. (Note that the matrix P_θ is simply $(\theta_i \theta_j) = \theta\theta'$, if θ is written as a column vector with entries $\theta_1, \ldots, \theta_n$.)

Now $E_2' = T_2 T_2^t = T_1(c_1 I + c_2 P_\theta)^2 T_1 = c_1^2 T_1^2 + (2c_1 c_2 + c_2^2) T_1 P_\theta T_1$. In the case at hand, $\theta = s\|T_1 a\|^{-1} T_1 a$. It follows that

$$P_\theta = \|T_1 a\|^{-2}(T_1 a)(T_1 a)^t = \|T_1 a\|^{-2}(T_1)(a)(a)^t(T_1),$$

so

$$\begin{aligned} T_1 P_\theta T_1 &= \|T_1 a\|^{-2}(T_1^2)(a)(a)^t(T_1)^2 \\ &= \|T_1 a\|^{-2}(T_1^2 a)(T_1^2 a)^t = (E_1' a, a)^{-1}(E_1' a)(E_1' a)^t; \end{aligned}$$

i.e.,

$$E_2' = c_1 E_1' + (2c_1 c_2 + c_2^2)(E_1' a,\ a)^{-1}(E_1' a)(E_1' a)^t.$$

It remains to calculate d. But it is evident from affine geometry that d can be very simply calculated as follows: Let D be the distance from the origin to the point of intersection of the hyperplane H with the line determined by a. Let $D' = \|T_1 a\|^{-1}\|T_1^2 a\| = (E_1' a, a)^{-1/2}\|E_1' a\|$. Then $d = D/D'$.

We now describe the algorithm and then the proposed modification. The algorithm is this. Define a sequence $E_1, E_2, \ldots$ of ellipsoids as follows: Set $E_1 = B_o$. As has been remarked, if (*) is consistent, then E_1 must contain a piece S of the solution set having volume at least V_o. Assuming for the moment that (*) is consistent, we take as our first trial solution the

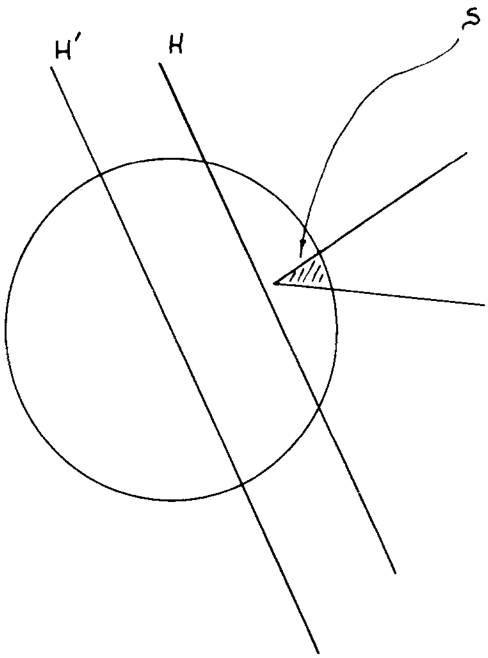

FIGURE 2

origin. If the origin satisfies (*) we are done. If not, there exists an inequality in (*) which is violated at the origin. Take the hyperplane H corresponding to this inequality, and parallel translate it to the center of E_1. Call the result H'. (See Fig. 2.)

Let $\frac{1}{2}E_1$ be the intersection of E_1 with the half-space determined by H' which contains S. By a previous argument, there exists an ellipsoid E_2 containing $\frac{1}{2}E_1$, and such that $\text{vol}(E_2) \leq r_n \text{ vol}(E_1)$ (note that E_2 contains S). Take the center of E_2 as the next trial solution. If it satisfies (*) stop. If not, there is an inequality in (*) which is violated at the center of E_2. If this is the case, repeat the just-described process. In this way, one constructs a sequence $E_1, E_2, \ldots$ of ellipsoids, each one of which contains S, and the volumes of which are decreasing exponentially. Ultimately this will come into conflict with the fact that S has volume greater than V_0, so a solution must be obtained before this happens. The number of repetitions of the process required to achieve this for a fixed n will be $O(\log R)$, where $R = (\text{vol}(B_0))/V_0$, and this estimate is polynomial in n and L. Since $r_n \to 1$ as $n \to \infty$, the adjusting constant implicit in the estimate will tend to infinity as $n \to \infty$. On the other hand, since $r_n^n < e^{-1/2}$, it is evident that there exists a constant C, such that at the very worst the number of repetitions is bounded by $Cn \log R$, which is also a polynomial estimate in

n and L. Finally, note that if (*) is not consistent, this will be detected by the algorithm, since if it has not obtained a solution after the above number of repetitions, then there cannot be a solution.

We conclude with the modification, which at this point can be stated very quickly. It seems quite clear that it may be very inefficient to repeatedly translate H back to H', and so always deal with half-ellipsoids. As we have seen, very little additional computational complexity is introduced by circumscribing instead the intersection of E_j with the "good" side of H. Computationally, the modified algorithm is the same as the original algorithm, with the change that c_1 and c_2 now involve d, and hence, are changing at each step. Additionally, it should be worthwhile to search for a distant H at each stage. Clearly, the estimated running time of the modified algorithm will depend on the accumulated products of the ratios $(\text{vol } E_{i+1})/(\text{vol } E_i)$, but can in no case be longer than the estimated running time of the original algorithm. Finally, we remark that recently several additional modifications of an empirical character have been proposed to accelerate the procedure, and as experience accumulates, some of these may be incorporated into various working versions of the algorithm (cf., e.g., [4, 6, and 7]).

REFERENCES

1. G. B. DANTZIG, "Linear Programming and Extensions," Princeton Univ. Press, Princeton, N.J., 1963.
2. P. GÁCS AND L. LOVÁSZ, Khachian's algorithm for linear programming, preprint, Computer Science Dept., Stanford University, Stanford, Calif.
3. L. G. KHACHIAN, A polynomial algorithm in linear programming, *Dokl. Akad. Nauk SSSR* **244**, No. 5 (1979), 1093–1096. [Translated in *Soviet Math. Dokl.* **20** (1979), 191–194.]
4. P. F. PICKEL, Some improvements to Khachian's algorithm in linear programming, preprint, Polytechnic Institute of New York, Farmingdale, N.Y.
5. N. Z. SHOR, Cut-off method with space extension in convex programming problems, *Kibernetika (Kiev)* 1 (1977), 94–95. [Translated in *Cybernetics* **13** (1977), 94–96.]
6. DONALD GOLDFARB AND MICHAEL J. TODD, Modifications and implementation of the Shor–Khachian algorithm for linear programming, preprint, Computer Science Dept., Cornell University, Ithaca, N.Y.
7. PHILLIP C. JONES AND EARL S. MARWIL, A dimensional reduction variant of Khachian's algorithm for linear programming problems, preprint, EG & G Idaho, Inc., Idaho Falls, Idaho.