Reconstructing Sets From Interpoint Distances

Paul Lemke
Steven S. Skiena
Warren D. Smith

Abstract

Which point sets realize a given distance multiset? Interesting cases include the "turnpike problem" where the points lie on a line, the "beltway problem" where the points lie on a loop, and multidimensional versions. We are interested both in the algorithmic problem of determining such point sets for a given collection of distances and the combinatorial problem of finding bounds on the maximum number of different solutions. These problems have applications in genetics and crystallography.

We give an extensive survey and bibliography in an effort to connect the independent efforts of previous researchers, and present many new results. In particular, we give improved combinatorial bounds for the turnpike and beltway problems. We present a pseudo-polynomial time algorithm as well as a practical $\mathbf{O}(2^n n \log n)$-time algorithm that find all solutions to the turnpike problem, and show that certain other variants of the problem are NP-hard. We conclude with a list of open problems.

1 Introduction

A set of n points in some space defines a set of distances between all pairs of points. In this paper we consider the inverse problem of constructing all point sets which realize a given distance multiset. The complexity of an algorithm to generate all such point sets depends upon the number of solutions, and so we are also interested in bounds on the maximum number of distinct solutions in a given space, as a function of n.

The problem dates back to the origins of X-ray crystallography in the 1930's [patt35] [picc39] [patt44]. More recently it has arisen in restriction site mapping of DNA, and was independently posed by M.I. Shamos [sham77] as a computational geometry problem. We encourage the reader to consult the recent thesis by Dakic [daki00] for the most recent results, including efforts to apply semi-definite programming to the problem. Pandurangan and Ramesh [pand01] have recent work on a variant of our problem which assumes additional information.

B. Aronov et al. (eds.), *Discrete and Computational Geometry*

Spaces of particular interest include restricting the points to a line or a circular loop. The analogy of these points as exits on a road lead us to call these cases the "turnpike" and "beltway" problems, respectively. A turnpike problem instance consists of a multiset of $\binom{n}{2}$ distances; a beltway instance consists of a list of $(n-1)n$ distances.

It should be made clear that the correspondences between the distances and point pairs are *not* known and the entire difficulty of the reconstruction problem is to deduce such labeling information. If the labels are known, then given the $\binom{n}{2}$ labeled distances among n points in d-space, a suitable set of coordinates may be determined in $\mathbf{O}(n^2d)$ time. Let the nth point lie at $\vec{0}$ and let the coordinates of the $n-1$ nonzero points be given by the columns of a $d \times (n-1)$ upper triangular matrix A. Define B by $B = A^TA$. Then $B_{ij} = \frac{1}{2}(q_{0i} + q_{0j} - q_{ij})$ for $1 \leq i, jn$, and $B_{ii} = q_{0i}$ where q_{ij} is the squared distance between points i and j. We may solve for A in terms of B consecutively column by column in time $\mathbf{O}(dn)$ per column, for a total runtime $\mathbf{O}(dn^2)$. If $d = n$, this algorithm is called the "Cholesky factorization" [golu83]. Alternately, we may find the "eigendecomposition" of $B = Q^T\Lambda Q$ where Λ is the diagonal matrix of $n-1$ real eigenvalues of B (in decreasing order; only the first d can be nonzero) and the columns of Q are the eigenvectors of B [golu83]. Q has orthonormal rows and columns. Then $\Lambda^{1/2}Q$ has d nonzero rows. Its $n-1$ columns give coordinates for our $n-1$ points. This approach has numerical advantages in situations in which our distances are contaminated by noise or roundoff error, because, e.g. the best approximation, in the Frobenius norm, of a symmetric matrix by a rank-d positive definite symmetric matrix is precisely its eigendecomposition with all eigenvalues besides the d largest ones, artificially zeroed ([golu83]; this is the symmetric case of the SVD approximation theorem).

1.1 Notation

$\mathbf{Z}$ is the set of integers and $\mathbf{R}$ the set of real numbers, while $\mathbf{S}^d$ denotes the unit d-sphere $\{\vec{x} \in \mathbf{R}^{d+1}\colon\ |\vec{x}| = 1\}$. We will use $(\mathbf{R}/\mathbf{Z})^d$ to denote a *flat d-torus,* i.e. the d-cube $[0,1]^d$ with opposite faces equivalenced. In X-ray crystallography, one must reconstruct a point set from its vector-differences modulo some d-parallelipiped unit cell; by taking an affine transformation this parallelipiped may be transformed to $(\mathbf{R}/\mathbf{Z})^d$. Our asymptotic notation $\mathbf{O}()$, $o()$, $\theta()$, $\sim$ follows [knut76].

An algorithm is said to run in *pseudo-polynomial time* if it runs in time polynomial in the size of its input, when this input consists of integers written as unary numbers [gare78]. Similarly, a problem is *strongly NP-complete* if it is still NP-complete even if the input is required to consist of unary integers. For convenience, we adopt the "real RAM" [prep85] as our model of computation in this paper, although we have taken care not to abuse its excessive power. All of our lower bounds, NP-completeness proofs, and