A Polynomial Algorithm for a Constrained Travelling Salesman Problem

J. H. Rubinstein
Department of Mathematics
University of Melbourne
Parkville, VIC 3052, Australia

D. A. Thomas
Department of Electrical and Electronic Engineering
University of Melbourne
Parkville, VIC 3052, Australia

N. C. Wormald
Department of Mathematics
University of Melbourne
Parkville, VIC 3052, Australia

Abstract

We give a polynomial-time algorithm for finding a solution to the Travelling Salesman Problem when the points given are constrained to lie on a fixed set of smooth curves of finite length.

1 Introduction

If a salesman, starting from his home city, is to visit each city on a given list and then return home, it is plausible for him to select the order in which he visits the cities so that the total of the distances travelled in his tour is as small as possible. Let us assume he knows, for each pair of cities, the distance from one to the other. Then he has all the data necessary to find the minimum, but it is by no means obvious how to use this data in order to get the answer. The Travelling Salesman Problem (TSP) is one of the most prominent of the combinatorial optimisation problems. A comprehensive discussion of the NP-hard TSP can be found in [7]. The concepts of computational complexity theory generated much research during the 1970's. An impressive survey is presented by Garey and Johnson [5].

Of course the probable intractibility of the general TSP does not exclude the possibility of polynomial solution procedures for special cases of the problem. In fact, one of the most surprising aspects of computational complexity theory is that one can find sharp borders that separate the easy problems from the hard ones.

In a previous paper [10] we discussed the NP-hard Steiner problem, which asks for the shortest network connecting n points in the plane. We gave a polynomial-time algorithm for points constrained to lie on a fixed finite set of disjoint compact simple smooth curves in the plane.

In the Euclidean TSP the cities are given as *nodes* (points in the plane) and their distances are computed according to the Euclidean metric. (Hence the salesman is not restricted to travelling along any pre-existing roads. This would be the case if the salesman travels by helicopter, for instance.) The Euclidean TSP is NP-hard. However, suppose that the cities all lie on a small number of major highways (which we can assume are bounded and smooth). This should make the travelling salesman's problem easier. A similar instance of this problem would occur if a utility (such as optic fibre cable) had to to be connected to a set of sites on pre-existing roads, in a loop (i.e. with no branching). Motivated by this idea and our work in [10], we consider the following modification of the standard Euclidean TSP. Take a finite collection G of smooth compact curves. For any problem whose input is a set of points in the plane, restrict to the case that all input points lie on G. We call this a *G-constrained* problem. So for the G-constrained TSP the nodes must all lie on G. Our main result is Theorem 1, which gives a polynomial-time algorithm for this problem.

A series of related results has been obtained by Cutler [2], Rote [9] Deĭnekoet al. [3] and Deĭneko and Woeginger [4], giving polynomial algorithms for various families of Euclidean TSP instances. The family in [4] contains all the others, and consists of inputs in which the nodes not on the convex hull lie on k lines which intersect the hull in two common edges. These families are in some sense much more restricted than the ones considered in the present paper, and polynomials of quite low degree are achieved. However, the nature of these results is also different, since there is no pre-existing fixed set of curves (but see Note 5).

Theorem 1 *Let G be a finite set of smooth (i.e., continuously differentiable) compact curves with the following properties*

(i) there are a finite number of intersections, including self-intersections;

(ii) at any of these intersections the different branches of curves approach from different directions.

Then there is a polynomial-time algorithm for the G-constrained TSP.

Note 1. A compact curve must have finite length, and by the definition of continuously differentiable, the derivative must have a finite limit at the endpoints of the curve. This eliminates curve such as spirals and those like the graph of $x \sin \frac{1}{x}$. On the other hand, piecewise smooth curves are permitted, unless they violate condition (ii) of the theorem.

Note 2. We assume (unrealistically) that computing distances between given points can be performed precisely in bounded time, as can adding or comparing two such distances. This is partly to avoid the issue of how the curves are represented. It also avoids the difficulty that the

G-constrained TSP problem, as with Euclidean TSP, is not even known to be in NP without this assumption or some modification. This is due to the unknown complexity of deciding which of two polygonal paths is longer, even when the paths are made up of segments joining integer lattice points in the Euclidean plane; see [6] for example. The assumption may not sound practical, but in a practical algorithm, the computations will anyway involve rounding errors, and one will achieve a tour which is minimal only to within rounding errors. On the other hand, without this assumption, it is easily checked that our algorithm gives a tour of length within a factor $1+\epsilon$ of the minimum in time polynomial in the size of the input and in $1/\log\epsilon$ (assuming that the node coordinates are computable to k digit accuracy in time polynomial in k, which again brings in the issue of how the nodes are represented). Arora [1] recently gave a nearly linear-time approximation algorithm, and Rao and Smith [8] improved this to $n\log n$, but neither of these algorithms is polynomial in $1/\log\epsilon$, nor even in $1/\epsilon$.

Note 3. The nodes are the input for a given instance of the problem. So, in particular, in the statement of the theorem, by "polynomial-time" we mean the execution time is bounded above by a polynomial, which depends on G, in the number n of nodes. The degree resulting from our proof as given here is huge in general; the method would give a much lower degree if the choice of various constants were optimised (for instance, a much larger angle than 1° will suffice in Lemma 1 below), but then the presentation would be more complicated.

Note 4. We assume that the intersections (and self-intersections) of the curves are given in advance. We also assume that the input includes a partitioning of the nodes into sets along small sections of the curves which are "straight" enough (to be made precise in the proof). These two assumptions are not crucial: given any input G, in practice the intersections can be computed in a fixed time, as can the boundaries of the required small sections of curves, so that the required partitioning of nodes can be done in linear time. For instance, if the curves are given parametrically, then the ends of the sections of curve are given as values of these parameters, as are the nodes, and the partitioning is easy. The partitioning can also be induced by putting appropriate circles around the sections of the curves, so that the resulting closed discs determine the partition.

Note 5. In fact our algorithm can be phrased not in terms of curves at all, but in terms of a partition of the nodes existing as in Note 4. The feature the algorithm exploits is the bounded number of groups of nodes which are almost collinear within each group, as well as other properties where several groups come together (corresponding to an intersection of curves). We do not pursue this further here, but this would be a convenient framework if one desired the nodes to have rational coordinates.

A significant difference between the present result and the one in [10] for Steiner trees is that here for TSP, the curves in G can intersect at any non-zero angles, whereas in [10] the curves were permitted to form no angles unless close to 180°. This is a real difference between the two problems: it is shown in [10] that the Euclidean Steiner tree problem remains NP-hard even when the points are restricted to two line intervals meeting at an angle less than 120°, whereas for TSP this case has a polynomial-time algorithm.

We give an overview of the proof of the theorem, and then derive some properties of minimal tours (i.e. tours of minimal length) in Section 2. The algorithm is presented formally in Section 3.

Some terms used loosely here are defined precisely in the following section. For instance, we later define a term δ-straight to denote that a section of curve is "sufficiently" straight. The first two lemmas give conditions on the manner in which an optimal tour can pass through the nodes in a δ-straight section, such as not doubling back and forth too many times without leaving. Each node–node piece of a tour we will call an *edge*. These results are then used in Lemmas 3 and 4 to bound the number of distinct edges of a minimal tour that leave either a δ-straight section of a curve or a circle around an intersection, the bound being independent of the number of nodes.

With these preliminaries, the algorithm uses a dynamic programming approach. Circles are put around all the intersections of curves and the set of δ-straight sections of the curves corresponding to the (initial) partitioning of curves mentioned in Note 4. There is a fixed number of δ-straight sections and a bounded number of edges of the minimal tour can leave any one of them, there is a bounded number of edges of the tour between such sections, and hence a polynomial number of possibilities for these edges. These edges join pairs (a_i, b_i) of points where for each i, a_i and b_i are in different sections. Now the circles are shrunk, and when a node pops out of a circle we reach a different stage. For each possible pattern of connections (a_i, b_i) made by the tour between nodes in different circles, the minimum length way of achieving the pattern is computed using the information from the previous stage. Then the circles are shrunk and the process is repeated, and so on. Finally the circles have zero radius, so that the length of the connection pattern between different circles is equal to the length of the minimal tour. Along the way, we show that for each stage there are only a polynomial number of patterns to consider and the computation for each pattern is polynomial-time.

2 Properties of Minimal Tours

Suppose G is a finite set of smooth (i.e. continuously differentiable) compact curves. Then G has finite total length, and we take G fixed in the rest of the discussion. A set S of nodes on G is prescribed. A *tour* through S is a closed (possibly self-intersecting) curve passing through all nodes in S. Let L denote the length of a closed curve containing all of G. Then L is an upper bound on the length of a tour through S.

Let T be a minimal tour (i.e. tour of minimal length) through S. Note that each part of T between two consecutive nodes u and v will be a straight line, which we call the *edge* uv. Considering T as the image of an oriented circle under a continuous mapping, we define a *section* of T to be the image of a connected part of the circle. From this viewpoint we may consider T to be directed, thus inducing a circular ordering of the nodes of S. We use $T[x_1, x_k]$ to denote the section of T beginning at node x_1 and ending at x_k. Any section of non-zero length is oriented either positively or negatively depending on whether it is consistent with the orientation of T. Similarly, we define a section of G to be a (closed) subinterval of a curve Q of G, and for x and y lying on Q, the section of Q from x to y is denoted by $Q[x, y]$. We use $|Z|$ to denote the length of an object Z, be it a section, tour, or edge.

Let $0^\circ < \delta < 180^\circ$. (We keep the convention that δ is an angular measure.) If in a section of curve the directions of any two tangents to the curve differ by less than δ, we say that section of curve is *δ-straight*. By the assumptions on G, for any $\delta > 0$, there exists $l^*(\delta)$ such that any section of curve in G of length less than $l^*(\delta)$ is δ-straight. We also note the following, which is easily proved by calculus.

Proposition 1 *If $Q_0 = Q[x, y]$ is a δ-straight section of a curve Q then*

(i) any two straight line segments each having both endpoints on Q_0 have directions which differ by less than δ,

(ii) $|Q_0| < |xy|/\cos\delta$. ∎

As mentioned in the overview in Section 1, the strategy of the proof includes constructing small circles around intersections of the curves or self-intersections. These circles are made small enough such that the sections of the curves in the circles, acting as "spokes", are δ-straight for suitably small δ. The exact construction of these circles is discussed before Lemma 3. We begin with Lemmas 1 and 2, which will apply to sections of a curve which are sufficiently distant from all intersections.

The first lemma shows, roughly speaking, that T cannot "double back" too many times along a section of curve of G which is sufficiently straight.

Lemma 1 *Every 1°-straight section $Q_0 = Q[v, w]$ of a curve Q of G satisfies the following. Suppose that $T[v, w]$ contains no nodes of S other than nodes in $Q_0 = Q[v, w]$. Running in reverse along T from v, let u be the first node not in Q_0. Similarly, running forwards along T from w, let x be the first node which is not in Q_0. Then at least one of the angles uwv and wvx is more than 2° different from 180°.*

Proof. See Figure 1. Since Q_0 is 1°-straight, $|Q_0| \le 1.1|vw|$ by Proposition 1(ii). Let us assume that the conclusion of the lemma is false. Then without loss of generality the angle uwv is greater than 177° and so angle uvw is at most 3°. Thus since Q_0 is 1°-straight, $|T[u, v]| - |uw| \ge 0.9|vw|$. Similarly, $|T[w, x]| - |vx| \ge 0.9|vw|$.

Now obtain a new tour T' from T by replacing $T[u, x]$ by the edges uw and vx together with the curve Q_0. Then $|T| - |T'|$ can be computed as

$$(|T[u, v]| - |uw|) + (|T[w, x]| - |vx|) + (|T[v, w]| - |Q_0|).$$

Hence $|T| - |T'| \ge .9|vw| + .9|vw| - .1|vw| = 1.7|vw|$, contradicting the minimality of T. ∎

In view of Lemma 1, we shall define a *return* of T to be a node x of S such that the straight lines to the preceding and following nodes in T are at an angle of strictly less than 1°. Also, a *unidirectional* section of T is a section of T in which the nodes are in a δ-straight section of a curve of G and are visited in one of the two orders induced by orienting this curve.

The following lemma restricts the relative position of returns in a short section of G as long as there are no interfering curves nearby. In order to eliminate this interference we not only need different curves of G not to be too close, away from intersections, but that neighbouring

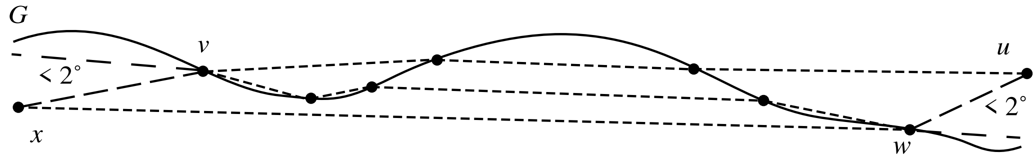

Figure 1: A forbidden configuration

δ-straight sections of a curve of G have similar directions. Before stating this lemma, it is convenient to select some constants which relate to the particular set of curves G.

First, we define any point p where curves in G intersect or self-intersect to be a *vertex* of *degree* k, where k is the number of different directions at which curves of G approach p. Let $V(G)$ denote the set of all vertices of G. Also, let $\delta_{\min}$ be smaller than the minimum angle which the tangents of two curves make at any vertex, and additionally smaller than $1°$. Choose $0 < r < \frac{1}{2}l^*(\frac{1}{4}\delta_{\min})$ such that the distance between any pair vertices is at least $3r$. By the assumptions on G, there is a constant $C_1 > 2000$ (C_1 sufficiently large) such that every point on a curve which is at least $C_1 l$ but less than r from any vertex has distance at least $2000l$ from any other curve of G, and all points on the same curve of distance less than $2000l$ from this point are in the same $1°$-straight section. Also by the properties of G, there is some minimum distance $d_{\min}$ between any pair of points which are both on curves of G, but not both in the same $1°$-straight section of one curve, and additionally lie at least r away from all vertices.

Now let l_0 be a positive constant such that $l_0 < \min\{r/2C_1, d_{\min}/1000\}$. Then any section Q_0 of a curve Q of G such that Q_0 has length $l < l_0$ and has distance at least $C_1 l$ from any vertex has the following property: there exists a $1°$-straight section $B(Q_0)$ of Q such that Q_0 has distance at least $1000l$ from $G \setminus B(Q_0)$.

Lemma 2 *There exists $l_1 > 0$ such that the following hold for any section Q_0 of a curve Q of G of length $l < l_1$ and of distance at least $C_1 l$ from any vertex in $V(G)$.*

(i) Let $T_1 = T[y, x]$ be a unidirectional positively oriented section of T with all its nodes in Q_0, and suppose that the next edge of T after x creates a return at x. Let $T_2 = T[x, u]$ be the maximal positively oriented unidirectional section of T starting with x and made up of nodes in $Q[x, y]$. Assume $u \neq x$. Let v be the next node visited by T after u. Then either v is in $Q[u, x]$ and there is a return at u, or v is not in $Q[y, x]$ and the angle vux is at most $1°$ different from $180°$.

(ii) There are at most two returns in any section of T all of whose nodes are in Q_0.

Note. It is difficult to strengthen part (ii) so as to eliminate the possible occurrence of two different returns in a short section Q_0, due to the possibility that the parts of T along Q_0

immediately before and after that section can be very short. To do this would require different hypotheses.

Proof. Define $l_1 = \min\{l^*(0.1°), l_0\}$ and take Q_0 of length $l < l_1$. In particular, this forces Q_0 to be 0.1°-straight. Moreover, by the statement after the definition of l_0 and the assumption about the distance from Q_0 to the nearest vertex, $B(Q_0)$ is defined, and is 1°-straight and contains all parts of G within $1000l$ of Q_0.
(i) The situations which must be shown to be the only possible ones are represented in Figure 2.

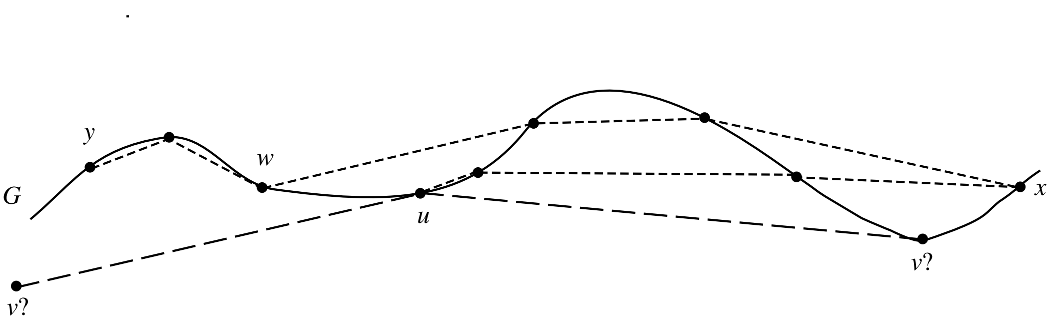

Figure 2: Two ways to go from u

If v is in $Q[u, x]$, then since Q_0 is 1°-straight, there is a return at u by Proposition 1(i).

We are left with the case that v is not in $Q[u, x]$. Then by maximality of T_2, v is not in $Q[y, x]$. All that remains to prove (i) is to show that the angle vux is at most 1° different from 180°. For this it is convenient to define w to be the last node visited by T_1 before arriving in $Q[u, x]$ (as in Figure 2), so that $w \in Q[y, x]$.

We take two cases.

Case 1. The edge uv has length less than $1000l$.
Here v must lie on $B(Q_0)$, which is 1°-straight. So either vux is less than 1° different from 180°, and we are finished, or it is less than 1°, in which case, as v is not in $Q[u, x]$, $T[y, v]$ must contain the configuration forbidden by Lemma 1 (where the roles of u, v, w and x in Lemma 1 are now played by the nodes w, x, u and v). This is a contradiction.

Case 2. The edge uv has length at least $1000l$.
Assume, contrary to what must be shown, that the angle vux is more than 1° different from 180°. Then $|ux| < \frac{1}{1000}|vu|$ and so

$$|vu| + |ux| \cos 0.5° > |vx|. \tag{1}$$

On the other hand, since $l_1 \leq l^*(0.1°)$, we have that $Q[y, x]$ is 0.1°-straight. Thus, using a little geometry,

$$|T[wx]| \geq |wx| \geq |wu| + |ux| \cos 0.2°, \tag{2}$$

regardless of the value of $|wu|$.

Obtain the tour T' by replacing $T[wv]$ by the edge wu, the curve $Q[u, x]$, and the edge xv. Then $|T[x, u]| \geq |xu|$, and $|Q[u, x]| \leq |ux|/\cos 0.2^\circ$ by Proposition 1(ii). Hence, from (1) and (2),

$$\begin{aligned} |T| - |T'| &\geq |T[wx]| + |T[x, u]| + |uv| - |wu| - |Q[u, x]| - |xv| \\ &\geq (\cos 0.2^\circ + 1 - 1/\cos 0.2^\circ - \cos 0.5^\circ)|ux| \\ &> 0.00002|ux|, \end{aligned}$$

contradicting the minimality of T, which completes Case 2 and thus the proof of (i).
(ii) Assume to the contrary that there are at least three returns in a section T_0 of T all of whose nodes are in Q_0. Since Q_0 is 1°-straight, by Proposition 1 every two distinct maximal unidirectional sections of T_0 are separated by a return. Let x and u be two consecutive returns in T_0 such that the $|xu|$ is minimised. Since there are at least three returns, we can assume without loss of generality that in T_0 there is a return previous to x, say y, so that $u \in Q[y, x]$.

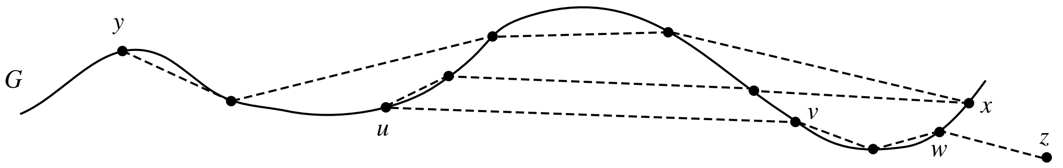

Figure 3: Angle zwu is approximately 180°

Let v be the next node after u in T_0, as in Figure 3. Since there is a return at u, it follows by (i) that $v \in Q[u, x]$. Let w denote the last node in the unidirectional section of T_0 in $Q[u, x]$ which begins with uv. By the assumption that x and u are the closest pair of returns, w cannot be a return. So, defining z to be the next node of T after w, z is not in $Q[y, w]$ because Q_0 is 1°-straight. Applying Lemma 1, with u, v, w and x standing for y, x, u and z, we deduce that one of the angles yux and uxz is at least 2° different from 180°. It cannot be the former, because Q_0 is 1°-straight, so it is the latter. Hence z cannot lie in $B(Q_0)$, and hence $|wz| > 1000l$. Given this, since uxz is at least 2° different from 180°, it follows that the angle uwz is more than 1° different from 180° (recalling that Q_0 is 0.1°-straight). However, applying part (i) with y, x, u and v standing for x, u, w and z, we obtain a contradiction to this statement. ∎

We next describe a partition of the curves of G into pieces small enough that the tour T cannot enter and leave the pieces more than a bounded number of times (regardless of the number of nodes). This is the key to obtaining a polynomial-time algorithm.

Let $\epsilon_0 > 0$ be chosen so that

(i) any two vertices have distance apart at least $3\epsilon_0$, and

(ii) $\epsilon_0 < \frac{1}{3}l^*(1^\circ)$,

(iii) each curve of G has length at least $3\epsilon_0$.

(The last condition is not necessary for the proof method, but makes the method easier to describe.) For $\gamma > 0$, let $C(p, \gamma)$ denote the circle of radius γ around a vertex p, and $\overline{C}(p, \gamma)$ the closed disc which it bounds. Define $l_2 = \frac{1}{4}\min\{l_1, \frac{\epsilon_0}{C_1}\}$ where l_1 is defined in Lemma 2. Recall that $C_1 > 3$ (and is actually much larger, but this is suitable for our purposes), so that $l_2 < \epsilon_0/12$, and note that the sections of curves of G outside all the circles $C(p, \epsilon_0)$ have length at least ϵ_0. Thus, there exists a finite set $\mathcal{Q}$ of simply connected compact sections of curves in G which overlap only on their end-points and with the properties that the closure of $G - \bigcup_{p \in V(G)} \bar{C}(p, \epsilon_0)$ is equal to the union $\bigcup_{Q \in \mathcal{Q}} Q$ and that for each section $Q \in \mathcal{Q}$, we have $\frac{1}{2}l_2 < |Q| < l_2$. Fix such a set $\mathcal{Q}$.

The bounds established in the following two lemmas depend on G.

Lemma 3 *For each subsection $Q_1 \subseteq Q \in \mathcal{Q}$, the number of distinct edges of T with one end in Q_1 and one end outside Q_1 is bounded above by a fixed number A_1, where A_1 depends only on G and not on Q or Q_1 or the number of nodes in S or their locations.*

Proof. For $Q \in \mathcal{Q}$ of length l, let Q_1 be a subsection of Q and consider an edge uv of T with $u \in Q_1$ and $v \notin Q_1$. We consider two cases. Firstly, uv may have length at least $\frac{1}{2}l$. Since $l > \frac{1}{2}l_2$, there are at most $4L/l_2$ such edges, where L is the upper bound on the length of T.

Secondly, uv may have length less than $\frac{1}{2}l$. Note that Q is distant at least $\frac{1}{2}\epsilon_0$ from any vertex. Define Q_0 to be the union of Q with both adjacent sections of curve of the same length l. Then by the definition of l_2, Q_0, of length $3l$, satisfies the hypotheses in Lemma 2 and in particular has distance at least $3000l$ from $G \setminus B(Q_0)$.

One implication of this and the 1°-straightness of Q is that $v \in Q_0$. Hence, uv is part of a maximal section T_0 of T with nodes in Q_0. If T_0 does not leave Q_0 at all, then it forms the whole of T. Otherwise, there are two possibilities. The first is that T_0 joins to some part of $B(Q_0) - Q_0$, and hence the section (including the joining edge) contributes at least $\frac{1}{2}l$ to the length of T. The second is that T_0 joins to $G \setminus B(Q_0)$, and hence the contribution is at least $3000l$. As $l > \frac{1}{2}l_2$, the total number of such sections containing edges uv of this type is bounded. Since each section has at most two returns in Q_0 by Lemma 2(ii), it can contain at most six edges with an end in Q_1 and the other not. So again, the number of such edges uv is bounded. ∎

Note. Even if A_1 did depend on Q, a common upper bound for all of them would exist anyway because there is only a finite number of different $Q \in \mathcal{Q}$.

We will need the following lemma for various values of γ. At first we take $\gamma = \epsilon_0$ but at later stages we take ever-smaller values.

Lemma 4 *Let $p \in V(G)$. For $\gamma \leq \epsilon_0$, the number of edges of T with one endpoint inside a circle $C(p, \gamma)$ around a vertex p and the other end outside or on that circle is bounded above by some number $A_2(p)$, where $A_2(p)$ is independent of the number of nodes in S and their locations but depends on G and p.*

Proof. Suppose the degree of p is k. Then by the definition of ϵ_0, there are precisely k sections of curves of G radiating out from p to $C(p, 2\gamma)$, which are all 1°-straight and we call *spokes.*

Let $T_{2\gamma}$ denote all parts of the minimal tour T in $\overline{C}(p, 2\gamma)$, and define $T'_{2\gamma}$ to be the following collection of curves:

(i) doubles of all the spokes, of approximate length $4k\gamma$ and certainly less than $5k\gamma$ because the spokes are 1°-straight;

(ii) a shortest-length perfect matching M of all the points $p_1, \ldots, p_{2r}$ at which segments of the tour enter $\overline{C}(p, 2\gamma)$ from outside (or leave to the outside);

(iii) one copy of $C(p, 2\gamma)$, of length $4\pi\gamma$.

To bound the length of M in (ii), note that taking alternate arcs of $C(p, 2\gamma)$ between the points $p_1, \ldots, p_{2r}$ results in a perfect matching of those points using arcs of total length at most $4\pi\gamma$, or at most $2\pi\gamma$ if we take the shorter of the two matchings obtainable in this way. M is even shorter than this because the arcs can be straightened.

Consider replacing $T_{2\gamma}$ by $T'_{2\gamma}$. This replacement gives a connected tour T' covering all the nodes in S, because the graph formed from $T'_{2\gamma}$ is a graph with all vertices of even degree except for the points $p_1, \ldots, p_{2r}$ on $C(p, 2\gamma)$. These vertices have odd degree, so that adding $T'_{2\gamma}$ to $T \setminus T_{2\gamma}$ gives all vertices even degree. The length of $T'_{2\gamma}$ provides an upper bound on the length L_γ of $T_{2\gamma}$:

$$L_\gamma < (5k + 6\pi)\gamma. \tag{3}$$

Now let uv be an edge of T with u inside $C(p, \gamma)$ and v not. Put $l = \frac{1}{2}\min\{l_1, \frac{\gamma}{2C_1}\}$ (with l_1 as in Lemma 2), and note that $l < \gamma/12$ as $C_1 > 3$.

Case 1. $|uv| \geq l/4$. For each such edge, the part inside $C(p, 2\gamma)$ has length at least $l/4$, so there are at most $4L_\gamma/l$ such edges.

Case 2. $|uv| < l/4$. Let Q_0 be the section of the spoke containing u such that Q_0 has length l and is centred at the point where the spoke meets $C(p, \gamma)$. Since spokes are 1°-straight, $u \in Q_0$. By the definition of l and ϵ_0, $B(Q_0)$ exists and furthermore Lemma 2 applies to Q_0. By the same argument as for short edges in the proof of Lemma 3, the maximal section T_0 of T containing u and with nodes in Q_0 contributes at least $l/4$ (which can easily be increased to $l/2$) to L_γ. So there are at most $4L_\gamma/l$ such sections, and by Lemma 2(ii), each has at most two returns. Such a section T_0 can cross $C(p, \gamma)$ at most three times using edges internal to T_0, and possibly twice more using the first and last edges, which may leave Q_0 at other angles and therefore not constitute returns. Hence at most $20L_\gamma/l$ such edges uv can arise in this way.

Putting the two cases together, there are at most $24L_\gamma/l$ edges of T of the specified type, which is a bounded number by (3) and the definition of l. ∎

3 The algorithm

Proof of Theorem 1. We use a dynamic programming approach. We assume that the input includes the set V of vertices and a partition of the n nodes into groups induced by $\mathcal{Q}$ as defined above Lemma 3. These determine ϵ_0, so that the circles $C(p, \epsilon_0)$ can be computed for $p \in V$. (If only the curves in G are given, computing these pieces of data requires some fixed amount of preprocessing time to find $\mathcal{Q}$, plus some fixed amount of computation per node to determine which member of $\mathcal{Q}$ or circle $C(p, \epsilon_0)$ contains it.) The only geometric operations which the rest of the algorithm require are the computation of distances between points and the comparison of sums of distances; the existence of the curves is then immaterial.

Let $\mathcal{C}_1$ denote the set of circles $\{C(p, \gamma) : p \in V(G)\}$, together with, for each $Q \in \mathcal{Q}$, a circle which intersects G just at the two endpoints of Q. The existence of the latter circles is ensured because of the properties of $\mathcal{Q}$. They are not crucial to the proof, but merely give a convenient way of specifying ever smaller subcurves of each Q. For $0 \leq \gamma \leq 1$, let $\mathcal{C}_\gamma$ denote the set of circles with the same centres as the circles in $\mathcal{C}_1$, but each with its radius multiplied by γ.

The discs defined by the circles in $\mathcal{C}_1$ are interior-disjoint and cover G. Hence for each input node z_i there is some unique γ_i, $0 \leq \gamma_i \leq 1$, such that z_i lies on one of the circles in $\mathcal{C}_{\gamma_i}$. By relabelling the nodes, we can assume without loss of generality that $\gamma_1 \geq \cdots \geq \gamma_n$. The algorithm will progress through discrete stages $1, \ldots, n$, corresponding to the γ_i. There is also an initial stage 0, defined so as to correspond to $\gamma_0 = 1$. When the nodes are in general position, the γ_i have n distinct values, and in the following exposition we assume this. If two happen to coincide it is easy to modify the argument appropriately by pretending that they are infinitesimally separated. This is simply done by regarding each node z_i as being interior to a circle in $\mathcal{C}_{\gamma_j}$ for $j < i$, even if it actually lies upon that circle. Thus, as γ decreases continuously from 1 down to 0, the nodes $z_1, \ldots, z_n$ emerge from the circles in $\mathcal{C}_\gamma$ one by one. The order in which they emerge with this definition is easily computed. By the assumption about general position, each node lies interior to some circle of $\mathcal{C}_1$. Hence, to summarise, for $0 \leq j \leq n$ the nodes $z_1, \ldots, z_{j-1}$ lie outside all the circles in $\mathcal{C}_{\gamma_j}$, z_j lies exactly on one of the circles, and $z_{j+1}, \ldots, z_n$ are all interior to circles in $\mathcal{C}_{\gamma_j}$.

Let T' be any closed tour with straight edges which visits all the nodes. Let $F = \{\{x_1, y_1\}, \ldots, \{x_k, y_k\}\}$ denote the set of edges of T' entering or leaving circles in $\mathcal{C}_{\gamma_j}$ in the sense that x_i lies inside one such circle and y_i lies outside or on that same circle. Then the maximal sections of T' on the nodes $\{z_1, \ldots, z_j\}$ have as endpoints the nodes in $\{y_1, \ldots, y_k\} \cap \{z_1, \ldots, z_j\}$ (i.e., all the y_i except those which lie inside a different circle from x_i). The pairs of y_i at ends of the same section induce a partition $\mathcal{P}$ of these nodes y_i into subsets of cardinality 1 or 2, where a singleton subset arises if the maximal section has no edges. Define the *state* s in stage j related to T' to be the ordered triple $(j, F, \mathcal{P})$ (where the cardinality of F may vary in different states in stage j). We also say T' is *compatible* with s, and write $s = s(T', j)$. For example, if T' is the tour shown in Figure 4(a), then $F = \{\{z_8, z_2\}, \{z_6, z_1\}, \{z_9, z_4\}, \{z_7, z_3\}\}$ and $\mathcal{P} = \{\{z_4, z_2\}, \{z_1, z_3\}\}$. Long dashes in the figure represent edges in F, short dashes the other edges in T'.

For each state s in stage j, let L_s denote the minimum, over all T' compatible with s, of the total length of the edges of T' on the nodes $\{z_1, \ldots, z_j\}$, i.e. which have neither endpoint interior to a circle in $\mathcal{C}_{\gamma_j}$. Also let E_s be an example of a set of edges achieving this minimum. The algorithm computes L_s and E_s for each state s in each stage j, for each $j = 0, 1, \ldots, n$.

Eventually, at stage n, there are no nodes lying inside the circles, and so there is only one

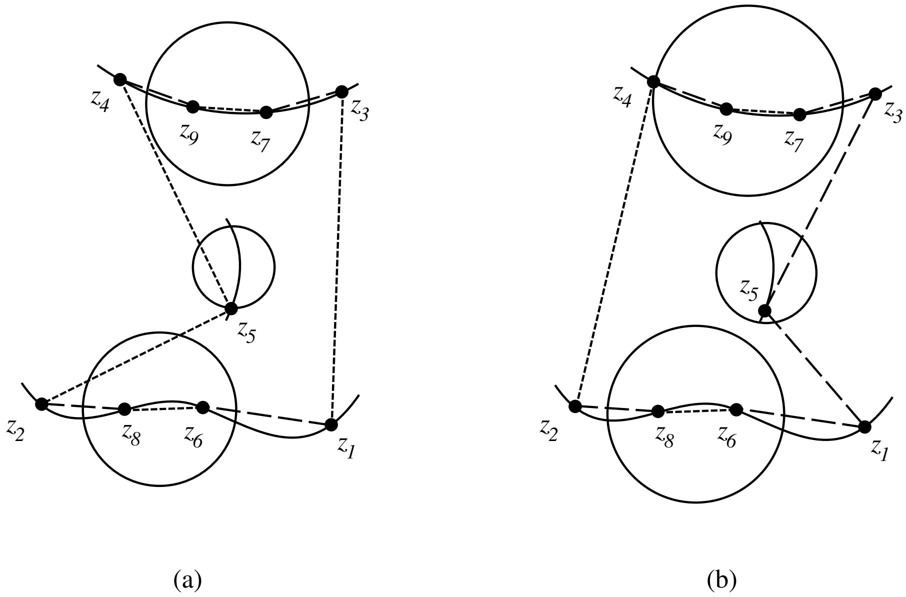

Figure 4: (a) A tour at stage 5. (b) A tour at stage 4.

possible state s, and furthermore L_s is the length of the minimal tour T and E_s is the edge set of such a tour. So the proof is completed by showing that for each stage there are only a polynomial number of states which need to be considered and that in each stage, for a state s, L_s and E_s can be computed in polynomial time from the data relating to the previous stage.

The minimal tour T must achieve the minimum in the definition of $L_{s(T,j)}$. Moreover, for $j \geq 1$ it will be evident from the scheme of the algorithm (and is part of the usual pattern of dynamic programming) that $L_{s(T,j)}$ and $E_{s(T,j)}$ are determined by $L_{s(T,j-1)}$ and $E_{s(T,j-1)}$. Hence, although it is not known during the algorithm which states are the $s(T,j)$, it is only these states for which L_s and E_s must be computed. Lemma 4 gives a bound on the number of edges of T leaving any circle in $\mathcal{C}_{\gamma_j}$ centred on a vertex of G, and Lemma 3 gives a similar bound for the other circles (letting Q_1 denote the section of a curve $Q \in \mathcal{Q}$ inside the circle). Thus, for the state $s(T,j)$,

$$|F| \leq A = qA_1 + \sum_{p \in V} A_2(p), \tag{4}$$

where $q = |\mathcal{Q}|$ and A_1 and $A_2(p)$ are the bounds in Lemmas 3 and 4. Any state s not satisfying this bound cannot be one of the $s(T,j)$ and therefore can be ignored (equivalent to putting $L_s = \infty$). This defines the states which are to be included in the algorithm at stage j: those in which (4) is satisfied and in which $\mathcal{P}$ is any possible pairing of $\{y_1, \ldots, y_k\} \cap \{z_1, \ldots, z_j\}$. For these states there is a polynomial number, at most n^{2A}, of possibilities for the set F of edges, and a bounded number of possibilities for $\mathcal{P}$ (given F). Hence, the number of states is bounded above by a polynomial.

We finish the proof by analysing the computation time. In stage 0 there is only one state

s, because there are no nodes outside all these circles $\mathcal{C}_0$. Thus $L_s = 0$ and E_s is empty. Next, consider any state $s = (j, F, \mathcal{P})$ in stage $j \geq 1$, let T' be a tour and let $s' = (j-1, F', \mathcal{P}')$ be the state $s(j-1, T')$. If T' achieves the minimum in the definition of L_s, then it must also in $L_{s'}$. It follows that

$$L_s = \min_{s'=(j-1,F',\mathcal{P}')\in\mathcal{S}} L_{s'} + L^+ \tag{5}$$

where $\mathcal{S}$ is the set of states $(j-1, F', \mathcal{P}')$ in stage $j-1$ for which $\mathcal{P}$ arises by modifying $\mathcal{P}'$ according to the edges in F' incident with z_j (if any), and L^+ is defined as the sum of the lengths of these edges. For example, consider the tour T' at stage 5 in Figure 4(a). At stage 4 the node z_5 is (was) inside its circle, and so the state $s(4, T')$, which is of course in $\mathcal{S}$, is equal to $(4, F', \mathcal{P}')$ where

$$F' = \{\{z_8, z_2\}, \{z_6, z_1\}, \{z_9, z_4\}, \{z_7, z_3\}, \{z_5, z_4\}, \{z_5, z_2\}\}$$

and $\mathcal{P}' = \{\{z_4\}, \{z_2\}, \{z_1, z_3\}$. Another state in $\mathcal{S}$ is the stage 4 state arising from the tour depicted in Figure 4(b), in which F' is the same but $\mathcal{P}' = \{\{z_4, z_2\}, \{z_1\}, \{z_3\}$.

The computation of (5) requires time $O(|\mathcal{S}|)$, which is polynomial since there is a polynomial number of states in stage $j-1$. (In fact, we can argue that $|\mathcal{S}| = O(n^2)$ using the fact that there are at most two edges in F' incident with z_j. But this argument is not necessary.) Theorem 1 follows. ∎

Acknowledgement We are thankful to the anonymous referee who gave very detailed comments on the deficiencies of an earlier version of this article as well as some helpful suggestions for improved presentation.

References

[1] S. Arora, Nearly linear time approximation schemes for Euclidean TSP and other geometric problems. In *Proceedings of the 38th annual Symposium on Foundations of Computer Science*, 1997.

[2] M. Cutler, Efficient special case algorithms for the N-planar traveling salesman problem, *Networks* **10** (1980), 183–195.

[3] V.G. Deĭneko, R. van Dal and G. Rote, The convex-hull-and-line traveling salesman problem: a solvable case, *Inform. Process. Lett.* **51** (1994), 141–148.

[4] V.G. Deĭneko and G.J. Woeginger, The convex-hull-and-k-line travelling salesman problem, *Inform. Process. Lett.* **59** (1996), 295–301.

[5] M.R. Garey and D.S. Johnson, *Computers and Intractability, A Guide to the Theory of NP-Completeness*, Freeman, San Francisco, 1979.

[6] D.S. Johnson and C.H. Papadimitriou, Computational Complexity, Chapter 4 in *The Travelling Salesman Problem* (E.L. Lawler et al., eds), Wiley, Chichester, 1985.

[7] E.L. Lawler, J.K. Lenstra, A.H.G. Rinnooy Kan and D.B. Shmoys (eds), *The Travelling Salesman Problem*, Wiley, Chichester, 1985.

[8] S.B. Rao and W.D. Smith, Improved approximation schemes for geometrical graphs via "spanners" and "banyans," NECI Tech. Report, December 1997.

[9] G. Rote, The N-line traveling salesman problem, *Networks* **22** (1992), 91–108.

[10] J.H. Rubinstein, D.A. Thomas and N.C. Wormald, Steiner trees for points constrained to curves, *SIAM J. of Discrete Math.* **10** (1997), 1–17.