CRITICAL POINTS FOR THE STEINER RATIO

By D.A. Thomas

Department of Electrical Engineering, University of Melbourne, Parkville, Victoria 3052, Australia

Abstract.

We find the critical points for the Steiner ratio function on the configuration spaces for 3 and 4 points.

§0 Introduction.

The Steiner problem is to determine a network S of shortest total length called a minimal Steiner tree which connects a given set of points $x_1, x_2, ..., x_n$ in the Euclidean plane $\mathbf{R}^2$. This has been shown to be an NP- complete problem [1]. There is an algorithm of Melzak [5] for finding S, but the number of steps increases exponentially with n.

Let T be a minimal spanning tree for the points, i.e. T is a shortest tree with vertices precisely at $x_1, x_2, ..., x_n$. (A Steiner tree is allowed to have additional vertices). If L_S, L_T denote the lengths of S, T respectively, then the Steiner ratio $\rho = L_S/L_T$. The Steiner ratio conjecture was introduced by Gilbert and Pollak [2] and asked if $\rho \geq \sqrt{3}/2$ for all sets of points. We were able to give new proofs of the cases $3 \leq n \leq 5$ in [8] and also established the conjecture for n = 6 [9] and for cocircular points [10]. We used the technique of Variational Calculus, the key idea being to perturb the position of the points $x_1, x_2, ..., x_n$. The conjecture was recently proved by Du and Hwang using the variational approach as a motivation.

As the Steiner problem is NP-complete many algorithms have been developed to find approximate solutions to the shortest network problem. One approach is to decompose the large problem into smaller problems consisting of subsets containing only a few points. An example of this approach is the T_1-network problem where full components (where the given points are

Australasian Journal of Combinatorics 13(1996), pp.259-282

all of degree one) consist of only one Steiner point. Another approach is to use an annealing algorithm where a shortest network is sought by allowing small increases in length in an intermediate approximation in order to move away from a possible locally minimal situation.

Recently A.A. Tuzhilin and A.O. Ivanov [11] have tackled the problem of classifying all the locally minimal networks with convex boundaries. They have solved the problem of describing the full (non degenerate) Steiner networks with convex boundaries in terms of polygonal triangulations.

In this paper the ratio function ρ is studied on the configuration spaces for $n = 3$ and $n = 4$. In particular *all* critical points for ρ are found, not just minima. As in classical theory one studies level sets to understand the nature of the function ρ and particularly the minimal trees for critical configurations. Where ρ has critical points the nature of the level sets changes. This knowledge will be useful for the various approaches to the Steiner problem. If the ratio is close to $\frac{\sqrt{3}}{2}$ then T_1 networks approximate the solution well. If the ratio is close to 1 a spanning tree is a good approximation. Note how close the ratio for the rhombus, $\frac{\sqrt{7}}{2}$ (.8819) is to the ratio for the equilateral triangle, $\frac{\sqrt{3}}{2}$ (.866). Yet the rhombus is a local minimum whereas the equilateral triangle is the absolute minimum.

§ 1 Index of critical points of the ratio function.

Definition S is called a *full* Steiner tree if it has 2n-2 vertices. Equivalently n of the vertices are $x_1, x_2, \ldots, x_n$ and the other n-2 vertices have three edges meeting at 120° angles. The latter points are called Steiner vertices (c.f. [2]).

A choice of S is called a *topology*. We parametrize the configuration $x_1, x_2, \ldots, x_n$ by the vector $Y = (y_1, y_2, \ldots, y_{2n-3})$ of lengths of the 2n-3 edges of S. By a homothety, we can assume if necessary that the sum of the coordinates of Y is one.

Definition The *configuration space*

$$\Delta = \{Y = (y_1, y_2, \ldots, y_{2n-3}) : \sum_{i=1}^{2n-3} y_i = 1,\ y_i \geq 0 \text{ for all } i\}.$$

Then Δ is a (2n-4)-dimensional simplex.

Let $T_1, T_2, \ldots, T_k$ be all possible spanning trees for $x_1, x_2, \ldots, x_n$. Define $\rho = L_S/\min L_{T_j}$. Note we are abusing notation by calling ρ the Steiner ratio, since we intend to keep the topology of S fixed. However if it can be shown that $\rho \geq \sqrt{3}/2$ on Δ for every choice of topology, then obviously the ratio conjecture will be established.

Clearly ρ is continuous and Gâteaux differentiable. In fact the directional derivative or differential of ρ in the direction of a vector v at Y in Δ is given by

$$D\rho(v) = \lim_{h \to 0} \frac{1}{h} (L_S(Y + hv)/L_{T_j}(Y + hv) - L_S(Y)/L_{T_j}(Y)),$$

where j is chosen so that $L_{T_j} = \min \{L_{T_1}, L_{T_2}, \ldots, L_{T_k}\}$ for all points $Y + hv$, for h sufficiently small. So T_j is the spanning tree which remains minimal along the line $Y + hv$. Clearly $D\rho(v)$ is continuous in Y and v; we do not specifically refer to Y as it is usually clear from the context.

Definition A *critical point* Y for ρ satisfies $D\rho(v) \geq 0$ for all v.

Lemma 1. $D\rho(v) = \dot{L}_T/L_T\,(\dot{L}_S/\dot{L}_T - \rho)$, where $T = T_j$ is given as above and $\dot{L}_T = DL_T(v)$, $\dot{L}_S = DL_S(v)$.

Corollary 1. If $\dot{L}_T < 0$ ($\dot{L}_T > 0$) then $D\rho(v) \geq 0$ is equivalent to $\dot{L}_S/\dot{L}_T \leq \rho$ ($\dot{L}_S/\dot{L}_T \geq \rho$ respectively).

Define $\rho_j = L_S/L_{T_j}$, for $1 \leq j \leq k$. Then each ρ_j is smooth and so we can compute its gradient

$$\nabla \rho_j = 1/L_{T_j} (\nabla L_S - \rho \nabla L_{T_j}).$$

Assume $T_1, T_2, \ldots, T_m$ are the minimal spanning trees for the configuration Y. We now give an analogous definition to that in Gromov [4].

Definition If $\nabla \rho_1, \nabla \rho_2, \ldots \nabla \rho_m$ lie in an open half-space of R^{2n-4} then Y is called a *regular point* of ρ.

Note Since Δ is a $(2n-4)$-simplex, its tangent space is of dimension $2n-4$. An open half-space is a component of the complement of a hyperplane.

Lemma 2. The critical points of ρ are precisely the non regular points.

Proof If Y is a regular point, choose a vector v orthogonal to the hyperplane and on the opposite side to the vectors $\nabla\rho_1, \dots \nabla\rho_m$. Then the inner product $< v, \nabla\rho_j> <0$ for all j. Hence Y is not a critical point, as $D\rho(v) < 0$. Conversely suppose Y is not a critical point. Consequently there is a vector v so that $D\rho(v) < 0$. But then $\rho_1, \rho_2, \dots, \rho_m$ are all decreasing in the direction of v, ie $< v, \nabla\rho_j > < 0$ for all j. So Y is a regular point, as all the gradients lie in a half-space with v as outward normal.

Remark In practice, it is often convenient to consider ρ as defined on the positive part of R^{2n-3}, i.e. to not divide out by homotheties. In this case it is obvious that since each ρ_j is constant along rays through the origin, $\nabla\rho_j$ lies orthogonal to the ray through Y.

Next we construct a smooth vector field on the set R of regular points in Δ, following [4]. As in lemma 2, if Y is a regular point then we can choose a vector v so that $<v, \nabla\rho_j> <0$ for $1 \le j \le m$. By continuity, the open set R can be covered by small balls B_α so that v can be extended to a (parallel) vector field v_α with $D\rho(v_\alpha) < 0$ on V_α. Choose a smooth partition of unity ϕ_α on R subordinate to the cover B_α. Hence $0 \le \phi_\alpha \le 1$, $\phi_\alpha = 0$ outside B_α and $\sum_\alpha \phi_\alpha = 1$ for all regular points. The desired vector field is given by $w = \sum_\alpha \phi_\alpha v_\alpha$. It is easy to see that $D\rho(w) < 0$ for all regular points and so ρ decreases along the integral curves of w.

Definition The *cone* C spanned by $\nabla\rho_1, \dots, \nabla\rho_m$ is defined by

$$C = \{\lambda_1 \nabla\rho_1 + \dots + \lambda_m \nabla\rho_m, \lambda_j \ge 0 \text{ for all } j\}.$$

Then Y is a regular point exactly when C is contained in an open half space and Y is a critical point is equivalent to C being a subspace of R^{2n-4}. So we obtain

Lemma 3. Y is a critical point if and only if there are $\lambda_j \ge 0$ for $1 \le j \le m$ with not all $\lambda_j = 0$ and $\lambda_1 \nabla\rho_1 + \dots + \lambda_m \nabla\rho_m = 0$.

If Y is a critical point let $C^{\perp}$ denote the orthogonal complement of C in R^{2n-4}, i.e. $C^{\perp}$ is the subspace of vectors perpendicular to C. Clearly $D\rho(v) > 0$ for v in C and $D\rho(v) = 0$ for v in $C^{\perp}$.

Let $D^2\rho(v)$ be the second derivative of L_S/L_T in the direction v, where $T = T_j$ is chosen so that $L_{T_j} = \min\{L_{T_1}, \ldots, L_{T_k}\}$ for Y + hv, with h small.

Lemma 4. $D^2\rho(v) = (\ddot{L}_S - \rho\ddot{L}_T)/L_T - 2\dot{L}_T\, D\rho(v)/L^2_T$, where $\ddot{L}_S = D^2 L_S(v)$, $\ddot{L}_T = D^2 L_T(v)$.

Corollary 2. If v is in $C^{\perp}$, then $D^2\rho(v) = (\ddot{L}_S - \rho\ddot{L}_T)/L_T$.

Definition The Hessian of ρ_j is the bilinear form $(v,w) \to D^2\rho_j(v,w)$. Let Hess(ρ_j) denote the restriction of this form to $C^{\perp}$ at some critical point Y.

If an inner product is chosen on $C^{\perp}$, then Hess(ρ_j) can be naturally identified with a self adjoint linear operator on $C^{\perp}$. Consequently, if E_j (respectively E_j') is the subspace spanned by the eigenvectors corresponding to the non-negative (resp. positive) eigenvalues, then $C^{\perp} \equiv E_j \oplus E_j'$ and E_j is orthogonal to E_j'.

Let E be the subspace of $C_k^{\perp}$ spanned by the union of the E_j and let $E' = \bigcap_{j=1}^{k} E_j'$. Clearly $C^{\perp} \equiv E \oplus E'$ and E is orthogonal to E′. Also $D^2\rho(v) \geq 0$ if v is in E and $D^2\rho(v) < 0$ if v is in E′.

Definition The index of a critical point Y is the dimension of the subspace E′.

Also Y is called a *non-degenerate* critical point if $D^2\rho(v) > 0$ for v in E.

Remark We find for n = 3 and n = 4, all the critical points are non degenerate. (See §2 and §3). We conjecture that this is true for arbitrary n.

The Poincaré-Hopf theorem states that if w is a smooth vector field on a closed manifold M with zeros $Y_1, Y_2, \ldots Y_N$ having indices $i_1, i_2, \ldots, i_N$ then

$$\sum_{j=1}^{N} (-1)^{i_j} = \chi(M),$$

the Euler characteristic of M. Suppose Y is a non-degenerate critical point for ρ. Then it is not difficult to check that the index of Y as a zero of the vector field w constructed previously, is exactly as required in this formula (see e.g. Milnor [6]). If M is not a closed manifold, then there is a boundary term in the formula. In the next section, we extend the configuration space Δ, for $n = 3$, to a 2-sphere to avoid this problem.

§2 Critical points in the case n = 3.

First, by the argument in §1 of [8], it follows that the only critical point in int Δ is the equilateral triangle, for the three point case. In fact it is shown there that for each configuration Y there is a direction v so that $D\rho(v) < 0$, except for $Y = (^1/_3, {}^1/_3, {}^1/_3)$.

We can include the 2-simplex Δ in a 2-sphere S^2, where $S^2 = \{(y_1, y_2, y_3) : |y_1| + |y_2| + |y_3| = 1\}$. Therefore S^2 is an octahedron in Euclidean 3-space. A point (y_1, y_2, y_3) with say $y_1 < 0$, $y_2 > 0$, $y_3 > 0$ is interpreted as a configuration as shown in Figure 1(b). Note that edges meet at 60° angles, unlike the usual Steiner networks.

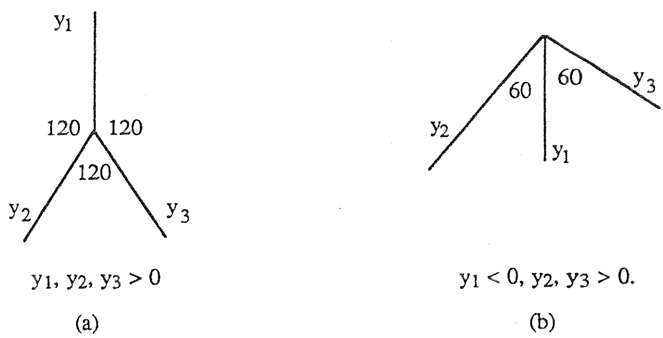
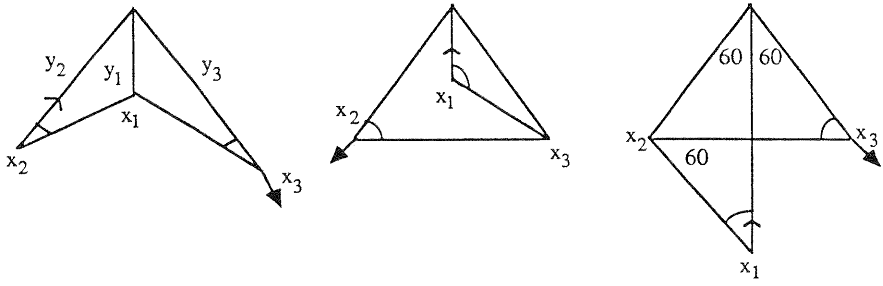


FIGURE 1.

The other possibilities for the signs of y_1, y_2, y_3 give diagrams as in Figure 1, with different labels on the edges. For example, $y_1, y_2, y_3 < 0$ corresponds to Figure 1(a).

Let us look for critical points Y in the interior of a 2-simplex Δ' in S^2 as in Figure 1(b), where $y_1 < 0$, $y_2, y_3 > 0$. Note that $\rho > 1$ occurs here. Suppose first that the points $x_1, x_2\ x_3$ are vertices of a non-isosceles triangle. There are 3 cases for T to consider, as shown in Figure 2.

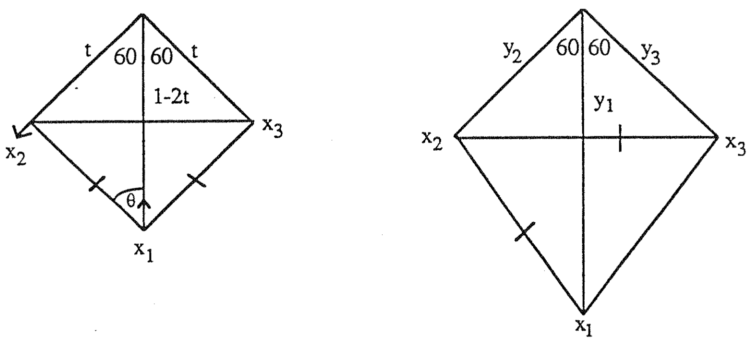


(a) (b) (c)

FIGURE 2

A variation is chosen as in Figure 2, with the properties that $\dot{L}_S = 0$ and $\dot{L}_T > 0$. Hence $D\rho\ (v) < 0$ and Y is not a critical point. Note in Figure 2(a), we are using the fact that the lengths $|x_1\ x_2| \neq |x_1\ x_3|$, so the angles at x_2 and x_3 are unequal. Similarly in Figure 2(b) it is easy to see the designated angles at x_1 and x_2 cant be the same. Finally in Figure 2(c), if the angles at x_1 and x_3 are equal, then angle $(x_3\ x_2\ x_1) = 60°$. But then $x_1\ x_3$ is not longest in the triangle $x_1\ x_2\ x_3$ so the tree $x_1\ x_2 \cup x_2\ x_3$ is not a minimal spanning tree.

There are 2 cases where $x_1\ x_2\ x_3$ is an isosceles triangle, as indicated in Figure 3.



(a) (b)

FIGURE 3.

In Figure 3(b), the cosine rule plus $|x_1 x_2| = |x_1 x_3|$ shows that $|y_1| = 1/2$, $y_2 + y_3 = 1/2$ and hence the triangle $x_1 x_2 x_3$ is equilateral. So we are reduced to Figure 3(a). Assume first that $|x_1 x_2| < |x_2 x_3|$. Then $L_T = 2\sqrt{1 - 5t + 7t^2}$, where $y_2 = y_3 = t$ and $|y_1| = 1 - 2t$. Differentiating with respect to t, we obtain a critical point when $t = 5/14$. Moreover ρ achieves its maximum value of $\sqrt{7/3}$. Also the index of this critical point is 2, since $T = x_1 x_2 \cup x_1 x_3$ is the *unique* minimal spanning tree for all configurations near $Y = (-4/14, 5/14, 5/14)$. Note that $\nabla\rho = (0, 0, 0)$ at this point, so Y is a standard (locally) *smooth* critical point for ρ.

If $|x_1 x_2| > |x_2 x_3|$, then we perform a variation as shown in Figure 3(a). We compute that $\dot{L}_T = -\cos 150 - \cos(60 + \theta) - \cos\theta$. This can be rewritten as $\sqrt{3}\left(\frac{1}{2} - \cos(\theta + 30)\right) > 0$, since $\theta > 30$. Hence $D\rho(v) < 0$ and the configuration is not a critical point.

It remains to consider the equilateral triangle, as shown in Figure 4.

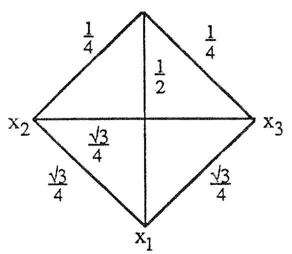


FIGURE 4.

In this case $\rho = 2/\sqrt{3}$. Let $T_1 = x_2 x_3 \cup x_1 x_3$, $T_2 = x_1 x_2 \cup x_2 x_3$, $T_3 = x_1 x_2 \cup x_1 x_3$. Then it is easy to verify that $\nabla\rho_1 = \nabla\rho_2 = (0, 0, 0)$, whilst $\nabla\rho_3 = (1, -1, 1)$. One might be tempted to conclude this is *not* a critical point, since the cone C spanned by these gradients is clearly a single ray, so lies in an open half-space.

However the analysis in §1 depended on knowing that $\nabla\rho_i$ is *non-zero*, which can easily be checked as correct for configurations where all the lengths are positive (*standard* Steiner trees). Here we must measure *second variation* of ρ_1 and ρ_2 in the direction $v = -\nabla\rho_3$. It is easy to check that if $D^2\rho_1(v) < 0$, then the configuration should *not* be regarded as a critical point, since the vector field w can be extended smoothly near Y. One can compute directly that $D^2\rho_1(v) < 0$ and so there is a single critical point in the interior of Δ'.

To complete the discussion, we investigate configurations lying on the edges of the octahedral 2-sphere. There are two types, where the non-zero coordinates have the same or opposite signs. It is easy to see that for points on the boundary of Δ, a vector v into Δ gives $D\rho(v) < 0$ and so there are no critical points of this type. The key situation is where one coordinate is positive, one is negative and the third is zero, as in Figure 5.

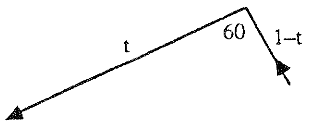
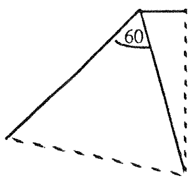


FIGURE 5.

Now $L_T = \min\{t, 1-t\} + \sqrt{3t^2 - 3t + 1}$. Differentiating, we obtain at most three critical points, namely $t = \frac{1}{2}$ and $t = \frac{1}{2} \pm \frac{1}{2}\sqrt{6}$. For $t = \frac{1}{2}$ it is rather easy to see directly that for any nearby configuration, $L_T < 1$.

We illustrate this in Figure 6. Hence $\rho > 1$ and as $\rho = 1$ at $t = \frac{1}{2}$, this is a local minimum.



$L_T < L_S = 1$

FIGURE 6.

For $t = \frac{1}{2} \pm \frac{1}{2}\sqrt{6}$, we know that $D\rho(v) = 0$ for v as in Figure 5. Hence $\nabla\rho_1$, $\nabla\rho_2$, $\nabla\rho_3$ must lie in the orthogonal directions to v. One easily checks then that Y is a saddle of index one for both values of t.

We summarize the results in Figure 7, where the octahedron has been projected onto the plane. The conclusion is that ρ has 8 minima, 12 saddles and 6 maxima. As expected by the Poincaré-Hopf theorem the alternating sum is 2, the Euler characteristic of the 2-sphere.

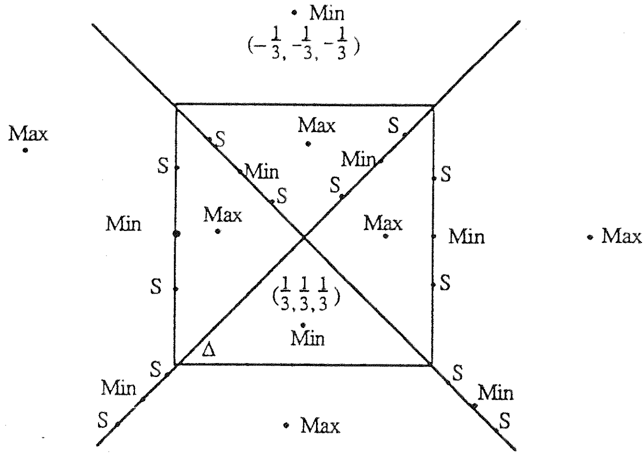


FIGURE 7.

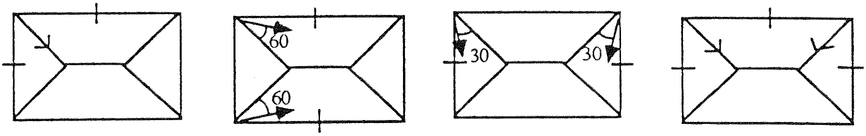
§3 Critical points for n = 4.

By Melzak's algorithm, it is not difficult to check that $\rho < 1$ for S a full Steiner tree, even though S may not be a minimal Steiner tree for the configuration Y of 4 points. To show that Y is not a critical point, it suffices to find a vector v so that $\dot{L}_T < 0$ and $\dot{L}_S/\dot{L}_T > \rho$ by Corollary 1. We will in fact usually show that $\dot{L}_S/\dot{L}_T \geq 1$.

As in [8], the union of all minimal spanning trees forms an embedded graph Γ. Also Γ is either a quadrilateral or two triangles which are isosceles or equilateral.

Case 1 Γ is a quadrilateral.

In Figure 8 we draw various possibilities for the equal (long) sides of Γ and indicate the variation vector v. In all cases, we leave it to the reader to verify that $\dot{L}_S/\dot{L}_T \geq 1$. Note that T always consists of Γ with a longest edge removed.



+ denotes long sides of equal length.

FIGURE 8.

The square and the remaining quadrilateral configuration with 3 equal long sides will be discussed later, as these give critical points (saddles) of index 1 and 2 respectively.

Case 2. Γ consists of two triangles.

Various situations for the equal long sides of Γ, with appropriate variation vectors v are given in Figure 9.

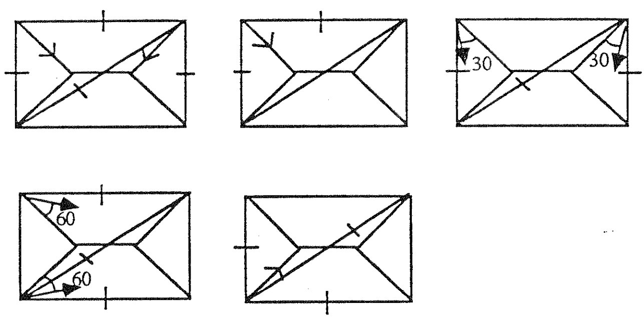


FIGURE 9.

Again, since $\dot{L}_S/\dot{L}_T \geq 1$ in all cases, it follows that there are no critical points in the pictures as in Figure 9.

Case 3 The square

We parametrize the edges of Γ and S for the square as in Figure 10. There are clearly 4 minimal spanning trees $T_i = \Gamma - a_i$, for $1 \leq i \leq 4$. So we have 4 ratios $\rho_i = L_S/L_{T_i}$ to consider. We choose a square of side length $\sqrt{3}$.

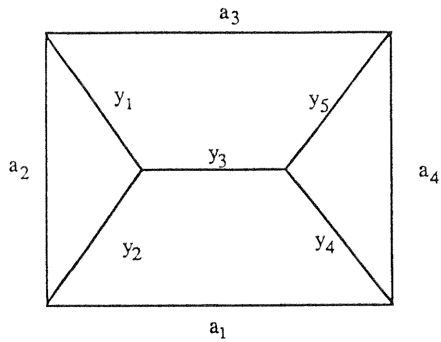


FIGURE 10.

To summarize the calculations, $\rho = (\sqrt{3} + 1)/3$, $\nabla L_S = (1, 1, 1, 1, 1)$ and homothety is in the direction $(1, 1, \sqrt{3} - 1, 1, 1)$. To compute ∇L_{T_i}, for example $\partial L_{T_i}/\partial y_1 = \cos 60° + \cos 30° = (1 + \sqrt{3})/2$. We calculate

$$6L_T \nabla\rho_1 = (2 - 2\sqrt{3}, 3 - \sqrt{3}, 4 - 2\sqrt{3}, 3 - \sqrt{3}, 2 - 2\sqrt{3})$$
$$6L_T \nabla\rho_2 = (5 - \sqrt{3}, 5 - \sqrt{3}, 2 - 4\sqrt{3}, 2 - 2\sqrt{3}, 2 - 2\sqrt{3})$$
$$6L_T \nabla\rho_3 = (3 - \sqrt{3}, 2 - 2\sqrt{3}, 4 - 2\sqrt{3}, 2 - 2\sqrt{3}, 3 - \sqrt{3})$$
$$6L_T \nabla\rho_4 = (2 - 2\sqrt{3}, 2 - 2\sqrt{3}, 2 - 4\sqrt{3}, 5 - \sqrt{3}, 5 - \sqrt{3})$$

These vectors span a subspace C of dimension 3 and in fact, $(\nabla\rho_1 + \nabla\rho_3)/(2 - \sqrt{3}) + (\nabla\rho_2 + \nabla\rho_4)/(2\sqrt{3} - 1) = 0$. Since all coefficients are positive, the square is a critical point. Also as all vectors are orthogonal to $v = (-1, 1, 0, -1, 1)$, we see that $C^\perp$ is the line along this vector.

Choose $y_1 = y_4 = 1 - t$, $y_2 = y_5 = 1 + t$, $y_3 = \sqrt{3} - 1$. Then clearly $L_S = \sqrt{3} + 3$ and $\dot{L}_S = \ddot{L}_S = 0$. Since $D\rho(v) = 0$, by Corollary 2, $D^2\rho(v) = -\rho\ddot{L}_T/L_T$. To determine the sign of $D^2\rho(v)$ we need to calculate $\ddot{L}_T$. By the cosine rule, $a_2 = a_4 = \sqrt{3 + t^2}$. Also $a_1 = a_3 = \sqrt{3t^2 + 3}$. Hence $L_T = 2\sqrt{3 + t^2} + \sqrt{3t^2 + 3}$, since $a_2 < a_1$ for $t > 0$. It follows that $\ddot{L}_T = 6(t^2 + 3)^{-3/2} + \sqrt{3}\,(t^2 + 1)^{-3/2}$. Consequently $\ddot{L}_T > 0$, $D^2\rho(v) < 0$ and the square is a critical point (saddle) of index 1.

Case 4 The rhombus.

Applying Melzak's algorithm, we see that $L_S = \sqrt{7}$, $L_T = 3$ and so $\rho = \sqrt{7}/3$.

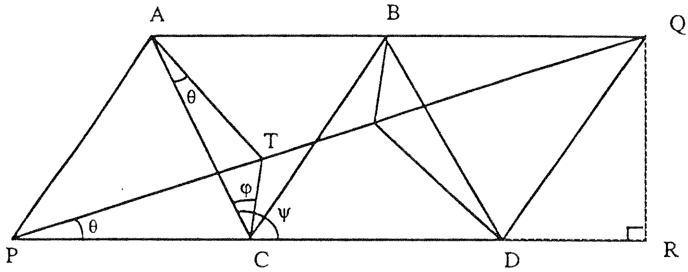


FIGURE 11.

In Figure 11, ABDC denotes the rhombus, so that ABC and BCD are equilateral. If $|AB| = 1$, then $|QR| = \sqrt{3}/2$, $|PR| = 5/2$ so $|PQ| = L_S = \sqrt{7}$, by Melzak. Let θ, φ, ψ be the angles as in Figure 11. Clearly $\varphi = 60 - \theta$, $\psi = 60 + \theta$. Also the angle QPR $= \theta$, since the circle through A, P, C also contains T and hence the segment TC subtends equal angles at A and P. Consequently $\cos\theta = 5/2\sqrt{7}$, $\sin\theta = \sqrt{3}/2\sqrt{7}$. It then follows that $\cos\varphi = 2/\sqrt{7}$, $\cos\psi = 1/2\sqrt{7}$.

Let a denote $\cos\theta$, $b = \cos\varphi$, $c = \cos\psi$. Then it can be readily checked that there are 8 minimal spanning trees for ABDC. We list ∇L_{T_i}, $1 \le i \le 8$ as follows:

(b, a, a + b, a, a + b + c) $T_1 = AB \cup BC \cup BD$
(b, a + c, 2a + b, b, a + c) $T_2 = AB \cup BC \cup CD$
(a + b, b, a, b + c, a) $T_3 = AC \cup AB \cup BD$
(a, b + c, a, b, a + b) $T_4 = AC \cup CD \cup BD$
(a, a + b + c, a + b, b, a) $T_5 = AC \cup BC \cup CD$
(a, a + b, b, a, a + b) $T_6 = AC \cup BC \cup BD$
(b, c, 2a, a + b, b + c) $T_7 = AB \cup BD \cup CD$
(a + b, b + c, 2a, b, c) $T_8 = AB \cup AC \cup CD$.

Now $\rho a = 5/6$, $\rho b = 2/3$, $\rho c = 1/6$ and $\nabla L_S = (1, 1, 1, 1,1)$. So we compute $L_T \nabla\rho_i = \nabla L_S - \rho\, \nabla L_{T_i}$ and multiply all entries by 6.

$$M = \begin{bmatrix} 2 & 1 & -3 & 1 & -4 \\ 2 & 0 & -8 & 2 & 0 \\ -3 & 2 & 1 & 1 & 1 \\ 1 & 1 & 1 & -3 & 2 \\ 1 & -4 & -3 & 2 & 1 \\ 1 & -3 & 2 & 1 & -3 \\ 2 & 5 & -4 & -3 & 1 \\ -3 & 1 & -4 & 2 & 5 \end{bmatrix}$$

Note that since we have computed $\sin\theta$ and $\sin\varphi$, the sine rule in the triangle ATC gives $|AT| = 2/\sqrt{7}$, $|CT| = 1/\sqrt{7}$. Hence the homothety is given by $(2, 1, 1, 2, 1)/\sqrt{7}$ .This is obviously orthogonal to all the rows in the above matrix M.

A row echelon form of M has non-zero rows given by:

$$\begin{bmatrix} 1 & 0 & -4 & 1 & 0 \\ 0 & 1 & 5 & -1 & -4 \\ 0 & 0 & -7 & 2 & 3 \\ 0 & 0 & 0 & 1 & -2 \end{bmatrix}$$

Consequently the gradients span the whole tangent space R^4 of the configuration space Δ.

Finally the reduced row echelon form of M^T is:

$$\begin{bmatrix} 1 & 0 & 0 & 0 & -1 & 0 & 1 & -1 \\ 0 & 2 & 0 & 0 & 1 & -1 & 1 & 2 \\ 0 & 0 & 1 & 0 & -1 & -1 & 1 & 1 \\ 0 & 0 & 0 & 1 & -1 & -1 & 2 & 0 \end{bmatrix}$$

Let λ_i, $1 \leq i \leq 8$, denote the coefficients of $\nabla\rho_i$ in the equation $\sum_{i=1}^{8} \lambda_i \nabla\rho_i = 0$. Then clearly $\lambda_3 = \lambda_4 = \lambda_6 = 1$, $\lambda_2 = 1/2$ and $\lambda_1 = \lambda_5 = \lambda_7 = \lambda_8 = 0$ is a solution which is non-negative. We conclude that $C = R^4$ and so the rhombus is a critical point of index 0, i.e. a local minimum.

Case 5 In Figures 8 and 9, there is one remaining case in each which turns out to give a critical point. One must resort to numerical calculations as there is only restricted symmetry. We consider Figure 9 first, as this case is easier than Figure 8.

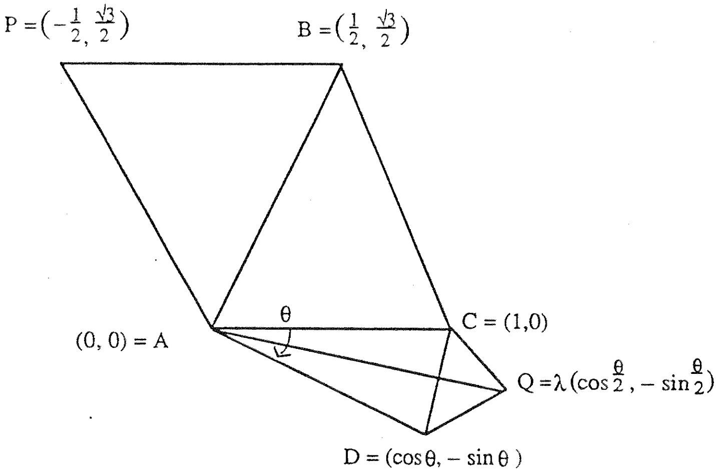


FIGURE 12

In Figure 12, ABC is equilateral and ACD is isosceles, with ABCD as the configuration. The idea is that as θ = angle CAD varies from 0 to 60, ABCD changes from a single equilateral triangle to a rhombus. Since both the latter figures are local minima, there should be an index one saddle between them by the "mountain pass" lemma.

Now Q is on the line through A at angle $\theta/2$ with AC. Clearly $\lambda = |AQ| = \cos\theta/2 + \sqrt{3}\sin\theta/2$, since $|CD| = 2\sin\theta/2$. Hence $\lambda = 2\cos(\theta/2 - 60)$. By Melzak, $L_S^2 = |PQ|^2 = (\lambda\cos\theta/2 + 1/2)^2 + (\lambda\sin\theta/2 + \sqrt{3}/2)^2 = 2\lambda^2 + 1 = 8\cos^2(\theta/2 - 60) + 1$ $= 4\cos(\theta - 120) + 5 = 4\sin(\theta - 30) + 5$.

Also $L_T = 2 + 2\sin\theta/2$.

Next we solve for $\dot{\rho} = (L_S/L_T)^{\dot{}} = 0$. This easily gives

$$4\cos(\theta - 30)\,(1 + \sin\theta/2) - \cos\theta/2\,(5 + 4\sin(\theta - 30)) = 0$$

So $4\cos(\theta - 30) - 5\cos\theta/2 + 4\sin(30 - \theta/2) = 0$. Hence $4(\sin(\theta + 60) + \sin(30 - \theta/2)) = 5\sin(90 + \theta/2)$ and $8\cos(3\theta/4 + 15)\sin(\theta/4 + 45) = 10\sin(\theta/4 + 45)\cos(\theta/4 + 45)$.

We conclude that

$$4\cos(3\theta/4 + 15) = 5\cos(\theta/4 + 45).$$

Calculating numerically, the solution is

$$\theta = 37.89°.$$

To complete this case, we need to show that the gradients corresponding to the 5 minimal spanning trees generate a 3-dimensional subspace C and that a non-negative linear combination of the gradients is zero. Then clearly $C^{\perp}$ will be in the direction v of θ varying and it is straight forward to check that $D^2\rho(v) < 0$.

First of all, since $L_S = \sqrt{5 + 4\sin(\theta - 30)}$ and $L_T = 2 + 2\sin\theta/2$, as $\theta = 37.89°$ we find $\rho = L_S/L_T = 0.8891$. Next the angles between the edges of S and T at the vertices need to be computed.

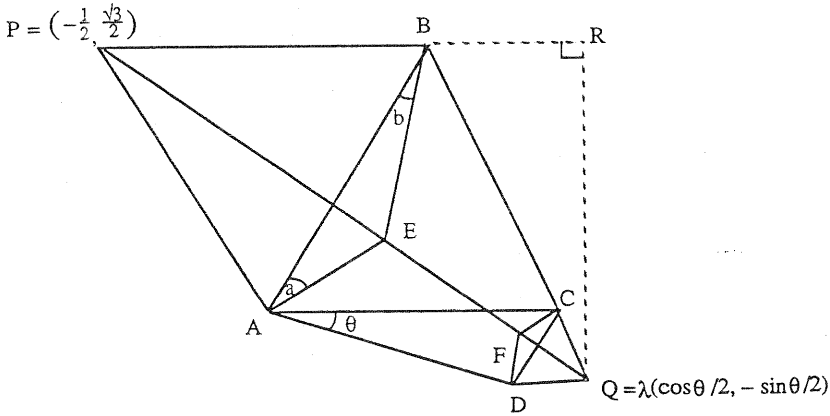


FIGURE 13.

As in Figure 13, since P, B, E, A lie on a circle centred at the barycentre of ABP, the angles BPE and BAE subtended by BE are equal. We denote these angles by a, similarly angle APE = angle ABE = b = 60 – a. Consequently,

$$\tan a = |QR|/|PR| = (\lambda \sin\theta/2 + \sqrt{3}/2)/(\lambda\cos\theta/2 + 1/2).$$

Because $\lambda = \cos\theta/2 + \sqrt{3}\sin\theta/2$, it follows that a = 35.13° and b = 24.87°. Note that angle

EAC = b = angle ACF. Since ACD is isosceles, angle ACD = 90 – θ/2 and hence angle DCF = 90 – b – θ/2. Also then angle FDC = b + θ/2 – 30 and all the angles needed for the calculation of the 5 gradient vectors are determined.

Let $T_1 = CD \cup AB \cup AD$, $T_2 = CD \cup AB \cup BC$,
$T_3 = CD \cup BC \cup AC$, $T_4 = CD \cup AB \cup AC$, $T_5 = CD \cup AD \cup BC$.

Then

$$L_T \nabla\rho_1 = (0.1933, -0.5338, 0.2729, -0.4222, 0.1366)$$
$$L_T \nabla\rho_2 = (-0.5338, 0.2729, 0.1933, 0.3049, 0.1366)$$
$$L_T \nabla\rho_3 = (0.2729, 0.1933, -0.5338, -0.5018, 0.1366)$$
$$L_T \nabla\rho_4 = (0.1933, -0.1341, 0.1120, 0.3845, -0.3445)$$
$$L_T \nabla\rho_5 = (0.2729, 0.5930, -0.6947, 0.3049, -0.3445)$$

These vectors span a 3 dimensional subspace of R^5. In fact a row echelon form is

$$\begin{pmatrix} 0.1933 & -0.5338 & 0.2729 & -0.4222 & 0.1366 \\ 0 & -1.2012 & 0.9469 & -0.8610 & 0.5138 \\ 0 & 0 & -0.1726 & -0.5845 & 0.3488 \end{pmatrix}$$

plus two rows of zeros, up to round-off error. To complete the argument that this configuration is critical, we can show that if $\lambda = (\lambda_1, \lambda_2, \lambda_3, \lambda_4, \lambda_5) = (0.6024, 1.0369, 0.8933, 1, 0)$, then $\sum_{i=1}^{5} \lambda_i L_T \nabla\rho_i = 0$, up to round-off error. Note that the *exact* solution with $\lambda_5 = 0$ will clearly satisfy $\lambda_i > 0$, for $1 \le i \le 4$.

Finally, since $L_S = \sqrt{5+4\sin(\theta-30)}$ and $L_T = 2 + 2\sin\theta/2$, we can readily compute $\ddot{\rho}$ in the direction of θ varying. As usual, $\ddot{\rho} = (\ddot{L}_S - \rho\ddot{L}_T)/L_T$. Also $\ddot{L}_S = -0.4168$, $\ddot{L}_T = -0.1624$ and $\rho = 0.8891$. Consequently $\ddot{\rho} = -0.1027$. This establishes that the configuration has index 1 as claimed.

Case 6.

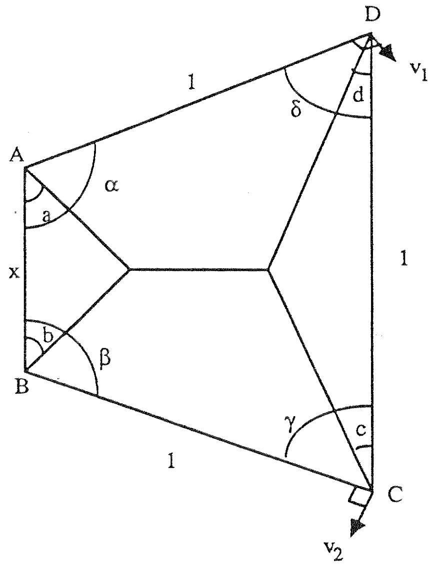


FIGURE 14.

In Figure 14, the configuration satisfies $|AB| = x < 1$, $|AD| = |DC| = |BC| = 1$, for suitable choice of scale. This is the final case corresponding to Figure 8. We call this configuration *symmetric* if $\alpha = \beta$. It is easy to see then that $\gamma = \delta$ and $\alpha + \delta = \beta + \gamma = 180$. The first problem is to eliminate the non-symmetric case.

Step 1   If ABCD is not symmetric then the configuration is not a critical point.

Without loss of generality, it can be assumed that $\delta > \gamma$ in Figure 14. We choose a variation v which fixes A, B and rotates DA about A and CB about B at speeds $|v_1|$ and $|v_2|$ respectively. To achieve $|DC|$ constant, v_1 and v_2 are chosen so that $|v_1| \sin\delta = |v_2| \sin\gamma$. Note then that $|v_1| < |v_2|$, supposing that $\delta + \gamma < 180$. The latter holds since $x < 1$.

We wish to compute the derivatives of all the angles a to δ, with respect to v. First, it is obvious that $\dot{\alpha} = -|v_1|$ and $\dot{\beta} = |v_2|$. Consequently, $(\dot{\alpha+\beta}) > 0$ and $\alpha + \beta$ is increasing. It is not difficult to see that the same variation v can be achieved by fixing DC and rotating DA and CB *upwards* (in Figure 14) at appropriate speeds. Hence $\dot{\delta} > 0$ and $\dot{\gamma} < 0$. We can choose v_1 and v_2 without loss of generality so that $\dot{\delta} = 1$. To find $\dot{\gamma}$, the equation linking γ and δ can be found. Introduce co-ordinates with C at the origin and D = (1, 0). Then B = $(\cos\gamma, \sin\gamma)$, A = $(1 - \cos\delta, \sin\delta)$ and

$$x^2 = |AB|^2 = (1 - \cos\delta - \cos\gamma)^2 + (\sin\gamma - \sin\delta)^2$$
$$= 3 + 2\cos(\gamma + \delta) - 2\cos\gamma - 2\cos\delta.$$

Differentiating relative to v, we obtain

$$-2\sin(\gamma + \delta)(1 + \dot{\gamma}) + 2\dot{\gamma}\sin\gamma + 2\sin\delta = 0.$$

Therefore $\dot{\gamma} = (\sin\gamma - \sin(\gamma + \delta)/(\sin(\gamma + \delta) - \sin\delta)$. Note that if we start at the symmetric configuration where $\gamma = \delta$, then $\dot{\gamma} = -1$. Obviously $\dot{\gamma} < 0$ remains true until $\sin\gamma = \sin(\gamma + \delta)$. Then $2\gamma + \delta = 180$ so a parallelogram can be constructed with all sides of length 1 and angles δ and 2γ, as in Figure 15. The appropriate diagonal gives two isosceles triangles with edges of length 1, 1, 1 + x and we conclude that $\beta = 180$!

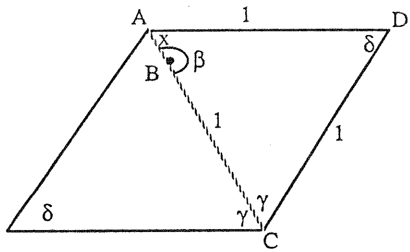


FIGURE 15.

We postpone till later the discussion when $\beta \geq 180$. So assume till further notice that $\beta < 180$ and $\dot{\gamma} < 0$, $\dot{\delta} = 1$. Note that $(\gamma \dot{+} \delta) = (\sin\gamma - \sin\delta)/(\sin(\gamma + \delta) - \sin\delta) < 0$, ie $\gamma + \delta$ is decreasing. Of course $\alpha + \beta + \gamma + \delta = 360$, so this is to be expected.

Next the behaviour of $\dot{a}$, $\dot{b}$, $\dot{c}$, $\dot{d}$ is described.

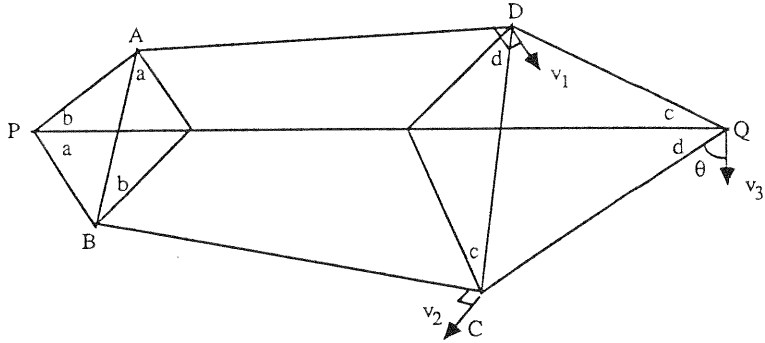


FIGURE 16.

Applying Melzak's algorithm gives Figure 16. As in Case 5, angle APQ = b etc. The vector v_3 gives the velocity of Q with respect to the variation v. To compute v_3, note that $|DQ| = |QC| = 1$ are constant. Hence

(+) $|v_3| \cos\theta = |v_2| \cos(\gamma - 30)$ and

$|v_3| \cos(120 - \theta) = |v_1| \cos(\delta - 30)$.

Combining these equations, we obtain

$$|v_1| \cos\theta \cos(\delta - 30) = |v_2| \cos(120 - \theta) \cos(\gamma - 30).$$

From before, $|v_1| \sin\delta = |v_2| \sin\gamma$ and so

(*) $\sin\gamma \cos\theta \cos(\delta - 30) = \sin\delta \cos(120 - \theta) \cos(\gamma - 30)$.

The initial equation (+) implies that $\theta < 90$ since γ is acute. Therefore PQ is turning "downwards" in Figure 16, ie $\dot{b} > 0$ and $\dot{a} < 0$. Similarly, by rewriting the variation so that A, B move and C, D are fixed, we see that $\dot{d} > 0$ and $\dot{c} < 0$. This completes the discussion of the variation of the angles in Figure 14.

Finally, since $L_S = |PQ|$ and $\dot{L}_T = 0$ with respect to v, a necessary and sufficient condition for $\dot{\rho} = 0$ is $\dot{L}_S = 0$, ie v_3 is orthogonal to PQ. In other words, if the configuration is critical then $\theta + d = 90$. Note that if $\dot{L}_S > 0$ relative to v, then $\dot{L}_S < 0$ with respect to $-$v. So if $\dot{L}_S \neq 0$ then the configuration is not critical.

Initially, $c = d = 30$ in the symmetric case and d is increasing. Consequently $d > 30$ and so $\theta < 60$ if $\theta + d = 90$ is to occur. But then equation (*) yields $\sin\gamma \cos(\delta - 30) > \sin\delta \cos(\gamma - 30)$. This implies $\cot\delta > \cot\gamma$, contradicting $\delta > \gamma$. So it follows that there are no non-symmetric critical points, in the range where $\beta < 180$.

If $\beta \geq 180$, then we claim that S is not a full Steiner tree, so this case doesn't occur.

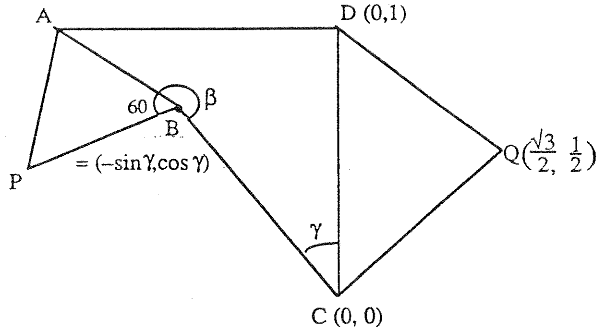


FIGURE 17.

It suffices to show in Figure 17 that angle PBQ > 180, ie slope PB > slope BQ. Now

$$\text{slope BQ} = (\cos\gamma - 1/2)/(-\sin\gamma - \sqrt{3}/2).$$

Also slope PB $> -\cot(\gamma + 60)$, since $\beta \geq 180$. If we set $-\cot(\gamma + 60) \geq (\cos\gamma - 1/2)(-\sin\gamma - \sqrt{3}/2)$, the inequality easily reduces to

$\cos(\gamma + 30) < \sqrt{3}/2$, which is trivially true. This completes step 1.

Step 2. Determination of critical configurations if ABCD is symmetric.

There is a one-parameter family of symmetric configurations, since $\alpha = \beta$, $\gamma = \delta$ and $\alpha + \gamma = 180$. Either x or γ serves as the parameter. From Figure 14, $\cos\gamma = (1 - x)/2$ or $x = 1 - 2\cos\gamma$. Applying Melzak,

$$L_S = \sin\gamma + \sqrt{3}/2(1 + x) = \sin\gamma + \sqrt{3}/2\,(2 - 2\cos\gamma)$$
$$= 2\sin(\gamma - 60) + \sqrt{3}.$$

Also $L_T = 2 + x = 3 - 2\cos\gamma$. We can readily solve $\dot{\rho} = 0$ with respect to γ varying. The result is $\sin(\gamma + 60) = 1/\sqrt{3}$ and so $\gamma = 84.74°$. (Note there are really two possible critical configurations, by taking |AB| = |AD| = |BC| = 1 and |CD| = x. This case will be discussed at the summary).

There is a one-parameter family of symmetric configurations, since $\alpha = \beta$, $\gamma = \delta$ and $\alpha + \gamma = 180$. In this case, $\cos\gamma = (1 - x)/2$, so equivalently $x = 1 - 2\cos\gamma$. Applying Melzak, we obtain $L_S = \sin\gamma + \sqrt{3}/2(1 + x) = \sin\gamma + \sqrt{3}/2(2 - 2\cos\gamma) = 2\sin(\gamma - 60) + \sqrt{3}$. Also $L_T = 2 + x = 3 - 2\cos\gamma$. The equation $\dot{\rho} = 0$ can be easily solved, differentiating relative to γ. The result is $\sin(\gamma + 60) = 1/\sqrt{3}$ and therefore $\gamma = 84.74$.

To show this is a critical point, we need to compute the 3 gradient vectors corresponding to $T_1 = AB \cup BC \cup CD$, $T_2 = AB \cup AD \cup BC$, $T_3 = AB \cup AD \cup CD$. Now

$$\nabla T_1 = (\sqrt{3}/2, \sqrt{3}/2 - \cos(\gamma + 30), \sin\gamma, \sqrt{3}/2 + \cos(\gamma - 30), \sqrt{3}/2)$$

$$\nabla T_2 = (\sqrt{3}/2 - \cos(\gamma+30), \sqrt{3}/2 - \cos(\gamma+30), 2\sin\gamma, \cos(\gamma-30) \cos(\gamma-30))$$

$$\nabla T_2 = (\sqrt{3}/2 - \cos(\gamma+30), \sqrt{3}/2, \sin\gamma, \sqrt{3}/2, \sqrt{3}/2 + \cos(\gamma - 30))$$

It is easy to check that $\lambda_1 \nabla\rho_1 + \lambda_2\nabla\rho_2 + \lambda_3 \nabla\rho_3 = 0$, where $\lambda_1 = \lambda_3 = 2\sin(\gamma + 30) - 1$, $\lambda_2 = 2 - 2\sin(\gamma + 30)$, using $\sin(\gamma + 60) = 1/\sqrt{3}$. Evaluating, we obtain $\lambda_1 = \lambda_3 = 0.8164$, $\lambda_2 = 0.1836$. Hence the configuration is critical and C is 2-dimensional.

Finally it remains to calculate the index. We observe first that $C^{\perp}$ is spanned by the vectors v as in step 1 (c.f. Figure 14) and w corresponding to γ varying. So it suffices to show $D^2\rho(v) < 0$ and $D^2\rho(w) < 0$. Relative to γ varying, we obtain $D^2\rho(w) = (\ddot{L}_S - \rho\ddot{L}_T)/L_T$, where $\ddot{L}_S = -2\sin(\gamma - 60)$, $\ddot{L}_T = 2\cos\gamma$, $L_S = 2\sin(\gamma - 60) + \sqrt{3}$, $L_T = 3 - 2\cos\gamma$ and $\gamma = 84.74$ at the critical configuration. The result is $D^2\rho(w) = -0.3553 < 0$.

For the variation v in Step 1, L_T remains constant so $\dot{L}_T = \ddot{L}_T = 0$. Therefore $D^2\rho(v) = \ddot{L}_S/L_T$. We introduce convenient co-ordinates into Figure 16 as follows. Let the origin be at the midpoint of the edge AB and assume PQ lies on the x-axis and AB on the y-axis. Then $A = (0, x/2)$, $B = (0, -x/2)$, $P = (-\sqrt{3}x/2, 0)$, $C = (\sin\gamma, -1/2)$, $D = (\sin\gamma, 1/2)$, $Q = (\sqrt{3}/2 + \sin\gamma, 0)$. Also the angle of slope of AD is $90 - \gamma$ and of BC is $\gamma - 90$. We perturb C to C′ and D to D′ by choosing slopes of AD′ to be $90 - \gamma - t$ and of BC′ to be $\gamma - 90 - t$. Consequently $D' = (\sin(\gamma + t), \cos(\gamma + t) + x/2)$ and $C' = (\sin(\gamma - t), -\cos(\gamma - t) - x/2)$. In particular the midpoint of C′D′ is $\sin\gamma(\cos t, \sin t)$ as expected.

We wish to find Q′ and it suffices to use Taylor expansions of second order, since the aim is to find $\ddot{L}_S$, where $L_S = |PQ'|$. The vector from C′ to D′ is $(\cos\gamma \sin t, 1 + 2\cos\gamma(\cos t - 1))$ $= (t\cos\gamma, 1 - t^2 \cos\gamma)$. Hence the orthogonal vector from the midpoint of C′D′ to Q′ is in direction $(-t^2 \cos\gamma + 1, -t\cos\gamma)$. The length of this vector is $1 - t^2 \cos\gamma + (t^2 \cos^2\gamma)/2$, up to second order. Therefore the co-ordinates of Q′ to second order terms in t^2 are given by adding $\sqrt{3}/2\ (1 + t^2(\cos\gamma - (\cos^2\gamma)/2)\ (1 - t^2 \cos\gamma, -t\cos\gamma)$ to the midpoint $\sin\gamma(\cos t, \sin t)$ of C′ D′. Simplifying and neglecting higher order expressions, we obtain

$$Q' = (\sqrt{3}/2 + \sin\gamma - t^2\ ((\sqrt{3}\cos^2\gamma)/2 + \sin\gamma)/2,\ t(\sin\gamma - (\sqrt{3}\cos\gamma)/2).$$

Finally, since $P = (-\sqrt{3}\ (1 - \cos\gamma), 0)$, $PQ' = (\sqrt{3} + 2\sin(\gamma - 60) - t^2\ ((\sqrt{3}\cos^2\gamma)/2 + \sin\gamma)/2,$ $t(\sin\gamma - (\sqrt{3}\cos\gamma)/2)$.

Consequently $L_S = |PQ'| = \sqrt{3} + 2\sin(\gamma - 60) + t^2((\sin\gamma - (\sqrt{3}\cos\gamma)/2)^2 - (\sqrt{3}\cos^2\gamma)/2 - \sin\gamma)$. The coefficient of t^2 in $2\ddot{L}_S = \sin\gamma(\sin\gamma - 1) + ((\sqrt{3}\cos^2\gamma)/2)\ (\sqrt{3}/2 - 1) - \sqrt{3}\sin\gamma\cos\gamma$ which is clearly negative. Hence the critical configuration has index 2.

SUMMARY

The interior of the 4-simplex, which is the configuration space in the case of 4 points, has the following critical points.

Number	Critical points	Index
1	Square	1
2	Rhombus	0
2	Equilateral and Isosceles	1
2	Symmetrical quadrilateral	2

The first column gives the number of such configurations. Note the action of the symmetry $A \leftrightarrow D$, $B \leftrightarrow C$.

It is interesting to observe that the Euler characteristic is one, which equals the sum

$$\sum_{\text{Critical points}} (-1)^{\text{index}}$$

Hence on the boundary of the configuration space, the negative gradient-like vector field w for ρ behaves as if it was everywhere inward-pointing, exactly as for the 3 point case. Of course this is only a statement about the *average* behaviour of w; near an equilateral triangle configuration w will be pointing outward towards it.

REFERENCES

[1] M.R. Garey, R.L. Graham and D.S. Johnson, *The complexity of computing Steiner minimal trees,* SIAM J. Appl. Math. 32, 1977, p.835-859.

[2] E.N. Gilbert and H.O. Pollak, *Steiner minimal trees,* SIAM J. Appl. Math. **16,** 1968, p.1-29.

[3] R.L. Graham, *Some results on Steiner minimal trees,* unpublished manuscript.

[4] M. Gromov, *Curvature, diameter and Betti numbers,* Comment. Math. Helv. **56,** 1981, p.179-195.

[5] Z.A. Melzak, *On the problem of Steiner,* Canad. Math. Bull. **4,** 1961, 143-148.

[6] J. Milnor, *Topology from the differentiable viewpoint*, University of Virginia Press, Charlottesville 1965.

[7] J. Milnor, *Morse theory,* Princeton University Press, Princeton 1968.

[8] J.H. Rubinstein and D.A. Thomas, *A variational approach to the Steiner network problem,* Annals of Operations Research, 33, (1991), 481-499.

[9] J.H. Rubinstein and D.A. Thomas, *The Steiner ratio conjecture for six points,* J. Comb. Theory Ser. A, **58** (1991) 54-77.

[10] J.H. Rubinstein and D.A. Thomas, *The Steiner ratio conjecture for cocircular points,* J. Discrete Comput. Geom. 7, 1992, p.77-86.

[11] A.A. Tuzhilin and A.O. Ivanov, *Local minimal networks with convex boundaries,* preprint.

(Received 22/5/95)