EXTREMAL ISOSYSTOLIC METRICS FOR COMPACT SURFACES

Eugenio CALABI

Mathematics Department
University of Pennsylvania
Philadelphia, PA 19104-6935 (USA)

Abstract. Given a closed, orientable surface M of genus ≥ 2, one seeks an *extremal isosystolic metric* on M: this is a Riemannian metric that induces on M the smallest possible area, subject to the constraint that the corresponding *systole*, or shortest length of any non-contractible closed curve, is a fixed, positive number. The geometric problem is rendered into an analytic one by reducing it to solving a nonlinear, partial differential equation with free boundaries. Examples are shown, to illustrate some possible candidates for solutions of the problem in special cases.

Résumé. Sur une surface M compacte orientable de genre ≥ 2, on cherche une métrique isosystolique extrémale : c'est une métrique riemannienne d'aire la plus petite possible sous la contrainte que la systole, i.e. la courbe fermée lisse non contractible de longueur minimale, soit un nombre positif fixé. Le problème géométrique est transformé en un problème analytique en le réduisant à la résolution d'une équation aux dérivées partielles non-linéaire à frontière libre. Des exemples sont donnés pour illustrer des candidats possibles à être solution du problème dans des cas particuliers.

M.S.C. Subject Classification Index (1991) : 53C22.

Supported by NSF Grant nr. 5-20600 during the preparation of this paper.

TABLE OF CONTENTS

1. INTRODUCTION

Given a compact Riemannian or Finslerian manifold (M, g), where g denotes the Riemannian (respectively, Finsler) metric, a base point $x_0 \in M$, and an element γ in the fundamental group $\pi_1(M, x_0)$, the *local systole* $Sys_\gamma(M, x_0, g)$ is defined to be the minimum length of any loop path through x_0 in the homotopy class γ. Denote by $\overline{\gamma}$ the conjugacy class of γ in $\pi_1(M, x_0)$; then the free local systole of (M, g) at $\overline{\gamma}$ is defined to be the minimum length of any closed path representing the free homotopy class $\overline{\gamma}$, and is denoted by $Sys_{\overline{\gamma}}(M, g) = Inf_{x_0 \in M}(Sys_\gamma(M, x_0, g))$. The systole (with no added qualifier) $Sys(M, g)$ is understood to be the least value of $Sys_{\overline{\gamma}}(M)$ as $\overline{\gamma}$ ranges over all non-trivial free homotopy classes.

In the terminology of M. Gromov [6], an n-dimensional, differentiable manifold M is called *essential*, if, for all Riemannian (respectively, Finsler) metrics g in M, the isosystolic ratio [1] $Vol(M, g)/(Sys(M, g))^n$ has a positive lower bound depending only on the topology of M. Gromov's compactness theorem asserts that, if M is essential, then for any positive constant c the function space of all metrics g in M, normalized by a positive factor so that $Sys(M, g) = 1$ and satisfying the volume inequality $Vol(M, g) \leq c$, is compact in the Fréchet-Hausdorff topology. In particular, all closed, 2-dimensional surfaces except for the 2-sphere are essential. With these facts in mind, it is natural to raise the question of estimating the minimum isosystolic ratio for any closed surface, orientable or not, in terms of its genus. Many variants of this question have been studied, some of them formulated to include more general spaces, such as manifolds with boundary, others dealing with restricted classes of metrics, such as Riemannian metrics with non-positive, or constant, negative curvature, or metrics in a given, conformal class, to name a few. While some statements in this paper apply to surfaces with boundary, we shall limit our consideration almost exclusively to Riemannian metrics in closed, orientable surfaces, leaving other cases for another occasion. The only types of closed surfaces for which one knows an explicit, extremal

[1] In Gromov's definition the isosystolic ratio is expressed by $Sys(M,g)/(Vol(M,g))^{1/n}$.

isosystolic metric, i.e. a Riemannian metric minimizing the isosystolic ratio, are the projective plane (P.M. Pu, [7]), the torus (C. Loewner, unpublished, cf. M. Berger, [3,4]) and the Klein Bottle (C. Bavard, [1,2]). For each of the other types of surfaces (i.e. for surfaces with negative Euler characteristic) there is a very wide gap between the best available estimates of upper and of lower bounds for the extremal isosystolic ratio. The main purpose of the paper is to reduce the problem of extremal isosystolic metric to a variational problem that may be studied by the methods of classical calculus of variations. At the end of this paper we shall exhibit for the record two explicit examples of metrics in an orientable surface of genus 3: both metrics attain locally minimum values of the isosystolic ratio, relative to small deformations of the metric in its function space, the second metric having an isosystolic ratio about 1.5% lower than the first ; it is believed that the value achieved by the second metric $((7\sqrt{3})/8 \approx 1,51554)$ is very close to, if not actually equal to the absolute minimum value for surfaces of genus 3. The two examples consist of piecewise flat metrics in the surface, each one constructed in terms of a corresponding, explicit, well known triangulation, with a large group of symmetries.

No similar construction has been found to yield an extremal isosystolic metric in surfaces of any genus $g = 2$, or ≥ 4, suggesting that the genera of surfaces whose extremal isosystolic metrics are piecewise flat may be quite sparse: it is this particular observation that has motivated the present study ; its ultimate goal is that of studying the general local properties of extremal isosystolic metrics, especially when they are not piecewise flat. Unfortunately the partial differential equations obtained have not yielded methods to construct any non-trivial, explicit solutions. However it is shown in Sections 6 and 7 that, merely by using the maximum principle, one can obtain some fairly close *a priori* estimates of the minimum isosystolic ratio in two examples, that illustrate also a useful generalization of the isosystolic problem. The first example consists of seeking a Riemannian metric in a 2-disk, admitting the group of symmetries of a regular hexagon, that minimizes the area subject to the condition that the least distance between each of the three pairs of opposite "sides" equals 2; the second example deals with the extremal isosystolic metrics in a torus with one open disc deleted: in this case the "systole" consists of two independent, positive, real numbers, representing, respectively, the "boundary systole" and the least length of any closed path representing a non-trivial homology class of cycles. Both of these examples illustrate some of the singularities that extremal isosystolic metrics may exhibit in general.

2. STRUCTURE OF k-REGULAR DOMAINS

Let M be a closed, orientable surface of genus $g \geq 2$ and consider the complete function space G of singular, generalized Riemannian (respectively, Finsler) metrics g on M, such that:

(i) g is bounded, locally, from above and below, by smooth Riemannian metrics ;

(ii) the g-length functional on the space of rectifiable arcs (the latter with the Fréchet topology) is lower semicontinuous.

This class of metrics is invariant under homeomorphisms of M of Lipschitz class; its definition ensures the compactness of any set of paths of bounded length, in any compact domain. In particular, the g-distance $d(x, y)$ between any two points $x, y \in M$ is achieved by a compact (non-empty) set of shortest paths. The function space G has the topology of uniform Lipschitz convergence of $d(x, y)$ in each compact subset of M: this topology ensures both the equivalence of the area functional $Vol(D) = Vol_g(D)$ [2] with the Lebesgue measure of any Borel set $D \subset M$ and its continuity with respect to the metric $g \in G$. Given any element $\overline{\gamma}$ in the set $\overline{\pi}_1^*(M)$ of non-trivial, homotopy classes of free, closed paths in M, the (free) local systole $Sys_{\overline{\gamma}}(M, g)$ is achieved by a compact family of oriented, closed paths of length $Sys_{\overline{\gamma}}(M, g)$, representing the class $\overline{\gamma}$: such closed paths will be referred to as <u>systole-long</u> paths ; for any given, positive real number A, the set $\Gamma_A \subset \overline{\pi}_1^*(M)$ consisting of all classes $\overline{\gamma}$ such that $Sys_{\overline{\gamma}}(M, g) \leq A$ is a finite set. The metrics in the class G may be discontinuous: for example, they may include isolated "short-cut" (or "fast-track") curves ; however it is a complete function space, to which an

[2] In the case of a Finsler metric g, the volume element form $dVol_g$ in terms of local parameters (u,v) is defined to be $\pi^{-1}\sigma(u,v)|du\wedge dv|$, where $\sigma(u,v)$ denotes the area of the unit g^*-disc in the cotangent bundle of M, with respect to the dual Finsler form g^* of g.

extremal isosystolic metric may be reasonably expected to belong, by a process of convergence of metrics whose isosystolic ratio approaches the lower bound, avoiding the full generality of the Fréchet-Hausdorff-Gromov topology.

Throughout the paper we shall tacitly assume that every metric $g \in G$ in M is normalized by the condition $Sys(M, g) = 1$, with possible exceptions explicitly stated.

Given any (M, g) with $g \in G$ and any real constant $A \geq 1$, consider the set $\Gamma_A \subset \overline{\pi}_1^*(M)$ consisting of all classes $\overline{\gamma}$ such that $Sys_{\overline{\gamma}}(M, g) \leq A$ and, for each $\overline{\gamma} \in \Gamma_A$, the union $K_{\overline{\gamma},A} \subset M$ of all oriented, systole-long paths representing the free homotopy class $\overline{\gamma}$. Let $\overline{\gamma}^{-1}$ denote the free homotopy class of the closed, oriented paths, whose reversal of orientation yields a path representing $\overline{\gamma}$: obviously $Sys_{\gamma^{-1}}(M, g) = Sys_{\gamma}(M, g)$, so that Γ_A is a finite, symmetric set, and $K_{\overline{\gamma}^{-1},A} = K_{\overline{\gamma},A}$. For any subset $S \subset \Gamma_A$ the subset $B_S = \cup_{\overline{\gamma} \in S}(K_{\overline{\gamma},A}) \subset M$ is compact, and therefore for any integer $k \geq 0$, the subset $U_{A,k} \subset M$ consisting of all points that are included in $K_{\overline{\gamma},A}$ for exactly $2k$ elements $\overline{\gamma} \in \overline{\pi}_1^*(M)$ (counting $\overline{\gamma}$ and $\overline{\gamma}^{-1}$ separately) is relatively open in the subset of points that are covered by at least $2k$ of the sets $K_{\overline{\gamma},A}$. For any $A \geq 1$, any non-empty, open subdomain $U \subset M$, contained in $U_{A,k}$ is called a <u>k-regular domain</u> in M.

Now assume that the metric g is an extremal isosystolic one ; we shall examine the possible open k-regular domains $U_{A,1} \subset M$ for small values of k.

Lemma 2.1. — *If g is an extremal isosystolic metric in M, then, for any constant $A \geq 1$, the subsets $U_{A,0}$ and $U_{A,1}$ of M are empty, and consequently $U_{A,2}$ is an open subdomain of M.*

Proof. The set $U_{A,0}$ is open. Hence, if it is not empty, there is a non-empty, open subdomain V such that its closure $\overline{V}$ is compact and $\subset U_{A,0}$. In addition there is a positive ϵ such that, for every point $x \in V$, the least length of any homotopically nontrivial loop based at x is $\geq A + \epsilon$. Replace the metric g by a conformally equivalent one $g' = g \cdot exp(-\delta\phi)$, where ϕ is a non-negative, non-zero function with support in V, and δ is a positive constant. Then the volume of M in terms of the metric g' is strictly smaller than the original one in terms of g ; at the same time, for δ sufficiently small, the systole $Sys(M, g')$ would remain identical with $Sys(M, g)$. This shows that

g could not be an extremal isosystolic metric, and consequently $U_{A,1}$ is an open subset of M.

Assume, as before, that there is a non-empty, simply connected, open set V such that $\overline{V} \subset U_{A,1}$ is compact, so that, for each point $x \in V$, there is at least one non-oriented, systole-long, closed curve passing through x ; any such curve with its two opposite orientations represents a unique, non-oriented, free homotopy class $\overline{\gamma}^{\pm 1} \in \overline{\pi}_1^*(M)$. While V is not necessarily foliated by its intersections with the covering family of systole-long paths, it is foliated by their orthogonal trajectories. From this foliation and a choice of orientation, say the one defined by $\overline{\gamma}$, one constructs a function u, that is constant along each orthogonal trajectory, and whose restriction to each of the systole-long paths provides its parametrization by its oriented arc length ; these properties determine u uniquely up to an added constant ; we shall refer to u as a potential function in V ; the formal definition follows this proof.

Given the potential function u in V, one chooses a second function v of Lipschitz class, such that (u, v) is a system of local parameters for M ; then the metric g may be represented almost everywhere as a quadratic form on the cotangent bundle, so that the norm $|\alpha|_g$ of a Pfaffian form $\alpha = \xi du + \eta dv$ is given by $|\alpha|_g^2 = \eta^2 + 2 \cdot f(u,v)dudv + g(u,v)\eta^2$ with $g(u,v) > (f(u,v))^2$ almost everywhere. Thus the metric is determined by the two functions $f(u,v)$, $g(u,v)$. The corresponding volume form is $dVol = (g(u,v) - (f(u,v))^2)^{-\frac{1}{2}}|du \wedge dv|$. As in the previous case, one could replace the metric g with another metric g', identical with g outside V and, inside V, defined by a quadratic form $|\xi du + \eta dv|_{g'}^2 = \xi^2 + 2 \cdot f(u,v)dudv + g'(u,v)\eta^2$ with $g'(u,v)$ slightly larger than $g(u,v)$ in a set of positive measure ; the resulting metric g' then would have an isosystolic ratio strictly smaller than that of g. This fact shows that, if g is an extremal isosystolic metric, then $U_{A,1}$ is empty as well as $U_{A,0}$. It follows that, for any extremal g and for any constant $A \geq 1$, $U_{A,2}$ is an open subdomain of M.

Definition 2.2. — *Given, in a surface (M, g), a family of oriented paths of shortest length, filling an open, simply connected domain $U \subset M$, and such that no two of the paths cross each other, a (local) geodesic potential function for the family of paths in U is a function $u : U \to \mathbf{R}$, that is constant along an orthogonal trajectory of the paths* *of the given family, and whose restriction to each path provides a parametrization of that path by its oriented arc length.*

It is well known that the choice of orthogonal trajectory in the above definition is immaterial, so that a local geodesic potential function for a given family of shortest paths is unique up to an additive constant. From a local viewpoint, a geodesic potential u is a function of Lipschitz class, with Lipschitz coefficient identically equal to 1 everywhere in V and no topological critical points (this second property must be added, since u is not necessarily differentiable) ; any function with these two properties has a gradient flow, whose orbits are shortest paths in V.

Lemma 2.3. — *If g is an extremal isosystolic metric in M, then the open subset $U_{A,2} \subset M$ is locally flat, and the systole-long paths belonging to the two distinct homotopy classes that meet at each point of $U_{A,2}$ intersect each other orthogonally almost everywhere.*

Proof. Let U be a simply connected subdomain of $U_{A,2}$, let $\overline{\gamma}_1, \overline{\gamma}_2$ be two distinct, non opposite homotopy classes of closed paths, among the four that are representable by systole-long paths of length $\leq A$, that meet U, and let u, v be potential functions in U, as in Definition 2.2, for the two respective families of systole-long paths in U. One deduces from Lemma 2.1 , after replacing U, if necessary, by a smaller domain, that the paths of the two families passing through any given point $x \in U$ cross each other transversally ; hence the pair of local potential functions (u, v) forms a system of local parameters. In any set where the angles between the paths are bounded away from zero, the Riemannian distance function, expressed in terms of (u, v) is of Lipschitz class ; it follows from Rademacher's theorem that (u, v) is differentiable almost everywhere in U. Therefore one may represent almost everywhere the Riemannian metric, as before, as a quadratic form on the cotangent bundle in U. The norm $|\alpha|_g$ of any given Pfaffian form $\alpha = \xi du + \eta dv$ in U is now defined by the quadratic form

$$|\alpha|_g^2 = |\xi du + \eta dv|_g^2 = \xi^2 + 2f(u,v)\xi\eta + \eta^2 \ , \tag{2.1}$$

where $f(u, v)$ is a measurable function satisfying $|f(u, v)| < 1$ almost everywhere, and the corresponding volume element is

$$dVol = (1 - (f(u,v))^2)^{-\frac{1}{2}} |du \wedge dv| \ \ ; \tag{2.2}$$

the statement of Lemma 2.2 is therefore reduced to showing that $f(u,v) = 0$ almost everywhere in U.

Suppose that $f(u,v)$ is not identically zero in a subset $V \subset U$ of positive measure. One could then alter the metric g in V, as in the proof of Lemma 2.1, by replacing $f(u,v)$ by its product with a smooth function slightly smaller than 1, so that the value of the systole of M with respect to the new metric would be unchanged ; at the same time the resulting total volume of M would be decreased. Therefore the given metric g could be an extremal isosystolic one in M, only if the local functions $f(u,v)$ defined in $U_{A,2}$ were identically zero almost everywhere. This concludes the proof of Lemma 2.2.

The lemma just proved does not exclude the possibility that $U_{A,2}$ may contain a set of measure zero of singular points.

The next lemma generalizes the last one in the case of domains $U_{A,k}$ with $k \geq 3$. In order to state it, we must recall the notion of generalized angle, adapted from A.D. Aleksandrov, between two paths of shortest length with a common point of origin x, when the metric may be singular and the paths may fail to be differentiable at x.

In the first place, even if the point x is not an isolated point of intersection, one may assume without loss of generality that the two paths do not cross each other (in the topological sense) anywhere else in a neighborhood of x. In fact, if they meet and cross at any point $y \neq x$, the segments between x and y along the two paths have obviously equal length ; if one then redefines the two paths by interchanging their traces along the segments between x and y, the new paths are again length minimizing and have "fewer" crossings, since they now meet at y without crossing each other. By applying Zorn's lemma, for any pair of shortest paths (or rays) issued from a common origin x, one may replace it with another pair of rays, respectively of equal lengths and jointly tracing the same continuum as the original pair, and not crossing each other anywhere in a neighborhood U of x ; then there exists a simply connected, compact neighborhood $V \subset U$ of x, such that the union Γ of the two rays splits V into two compact, pathwise connected subsets $V' \cup V$", with Γ as their common boundary.

Let $\delta_0 > 0$ be sufficiently small, so that the closed metric ball $B(x, \delta_0)$ of radius δ_0 and center x is contained in V, let V' be one of the two "halves" of V that are bounded by Γ and, for any $\delta (0 < \delta \leq \delta_0)$, let $E'(\delta) = V' \cap B(x, \delta)$. Then the angle spanned by V' is defined to be number $2 \cdot \lim_{\delta \to 0}(Vol(B(x, \delta))/\delta^2)$, if the limit exists, or, if the limit does not exist, the generalized angle is defined to be the set of accummulation values of $2 \cdot \lim_{\delta \to 0}(Vol(B(x, \delta))/\delta^2$. If the metric is continuous and non-degenerate at x, the angle just defined coincides with the elementary notion ; in the singular case, the sum of the two opposite angles spanned by two rays does not necessarily equal 2π. However, if x is an interior point of a path of least length, then the total angle spanned by the two resulting rays from x is necessarily $\geq 2\pi$.

Lemma 2.4. — *Let g be an extremal isosystolic metric in a closed, orientable surface M and let $U \subset M$ be any k-regular domain ($k \geq 2$). Introduce, for each point $x \in U$, a family of k unoriented, systole-long, closed curves passing through x, with the property that the $2k$ free homotopy classes, represented by each of the k curves with its two orientations, constitute a complete list of the $2k$ homotopy classes thus obtainable. Consider, locally at x, the corresponding family of $2k$ segments of these paths, originating at x (rays), ordered in their natural, counterclockwise cyclic order in terms of an orientation of M. Then, for almost all $x \in U$, the total of the angles at x from each of these $2k$ rays to the next equals 2π, the two angles formed by any two opposite pairs of paths are equal and, most importantly, the angle at x between any two consecutive paths is $\leq \pi/2$.*

Proof. In the first place we recall from the proof of Lemma 2.3 that, choosing as local coordinates the geodesic potential functions (u, v) corresponding to two of the families of systole-long paths in U, the coordinates are differentiable almost everywhere and the Riemannian metric form (2.1) is determined by the measurable function $f(u, v)$; f has the property that $|f(u, v)| < 1$ almost everywhere and $(1 - (f(u, v))^2)^{-\frac{1}{2}}$ is integrable. The function $f(u, v)$ represents almost everywhere in U the cosine of the interior angle between the oriented, systole-long paths chosen to define u and v: this proves the first two, more elementary, assertions. Suppose that the main conclusion failed: this would mean that, on choosing (u, v) corresponding to a consecutive pair among the $2k$ oriented, systole-long paths, the resulting function $f(u, v)$ would be strictly positive in a set $V \subset U$ of positive measure, and, in V, the remaining $k-2$ unoriented, systole-long paths would be representable as graphs of monotone decreasing functions, either expressing v in terms of u or vice-versa. The remainder of the proof would be similar to the corresponding arguments in Lemma 2.3. Choosing a bounded, positive function $\phi(u, v)$ with support in V, one could construct a metric g'_ϵ from g by replacing the function $f(u, v)$ by $f_\epsilon(u, v) = f(u, v) \cdot exp(-\epsilon \cdot \phi(u, v))$ for any constant $\epsilon > 0$. The change could only leave unchanged or decrease the norms of the geodesic potential functions of the original $2k$ families of systole long paths meeting V. As a result, the length of these paths could not decrease under the change of metric ; for ϵ sufficiently small, no additional systole-long paths, representing homotopy classes other than the $2k$ original ones could appear, and the total area of M would decrease. The combined effect would be that the substitute metric g'_ϵ would have an isosystolic ratio smaller than that of g, contrary to the assumption. This completes the proof of Lemma 2.4.

The proofs of the two last lemmas demonstrate the importance of the geodesic potential functions ; indeed they play an essential role in what follows. The first task is to extend the notion of these potential functions, so that they are defined, in some sense, globally, rather than just in the union of the respective systolic paths ; this is the topic of the next section.

3. SYSTOLIC BANDS AND POTENTIAL FUNCTIONS

Let M be a closed, oriented surface of genus $g \geq 2$, and let g be any Riemannian metric on M in the class G defined at the beginning of Section 2. Let $M^\sim$ denote the universal covering surface of M with $\Pi : M^\sim \to M$ denoting the covering map, choose an arbitrary point $x_0^\sim \in M^\sim$ as a base point, and let $x_0 = \Pi(x_0^\sim)$ be the corresponding base point for M. The group $\pi_1(M, x_0)$ operates freely by translation on $M^\sim$ and the covering map $\Pi : M^{\sim} \to M$ induces in $M^{\sim}$ the "pull-back" metric $g^{\sim} = \Pi^*(g)$, which is invariant under the action of $\pi_1(M, x_0)$. Thus, for any non-trivial element $\gamma \in \pi_1(M, x_0)$ and any $x^{\sim} \in M^{\sim}$ with $x = \Pi(x_0) \in M$, the Riemannian distance $d_{g^{\sim}}(x^{\sim}, \gamma(x^{\sim}))$ describes the least length of any x-based loop in M, representing the homotopy class $\psi^{-1}\varphi\psi$, where $\psi = \Pi(\psi^{\sim})$, $\psi^{\sim}$ is a path in $M^{\sim}$ from $x_0^{\sim}$ to $x^{\sim}$, and φ is any loop in M based at x_0, representing the homotopy class γ. Then the free local systole, $Sys_{\overline{\gamma}}(M, g)$, is the value of $Inf_{x^{\sim} \in M^{\sim}}(d(x^{\sim}, \gamma(x^{\sim})))$.

Given any non-trivial $\gamma \in \pi_1(M, x_0)$, choose any $x^{\sim} \in M^{\sim}$ that minimizes the distance $d(x^{\sim}, \gamma(x^{\sim})) = Sys_{\overline{\gamma}}(M, g)$ and any path $\varphi^{\sim}$ from $x^{\sim}$ to $\gamma(x^{\sim})$ that achieves that distance as its own length. It is not hard to verify that, since M is an orientable surface, for any integer n one has the identity $Sys_{\overline{\gamma}^n}(M, g) = |n| Sys_{\overline{\gamma}}(M, g)$; thus, if one takes the union of the following translations of that path, $U_{-\infty<n<\infty}(\gamma^n(\varphi^{\sim}))$, one obtains a complete path (i.e. a complete geodesic, in the smooth case), that achieves the minimum distance between any two of its points. To any non-trivial, cyclic subgroup $\langle\gamma\rangle \subset \pi_1(M, x_0)$ generated by γ, one associates the family Σ_γ of all complete, unbounded "geodesics" $U_{-\infty<n<\infty}(\gamma^n(\varphi^{\sim}))$ generated by all possible paths $\varphi^{\sim}$ of length $Sys_{\overline{\gamma}}(M, g)$, connecting any suitable $x^{\sim}$ with $\gamma(x^{\sim})$. One calls the paths of the family Σ_γ the systolic band directed by γ, and its trace $B_{\gamma^{\sim}} \subset M^{\sim}$ is the union of all the paths belonging to Σ_γ.

Given a non-trivial element $\gamma \in \pi_1(M, x_0)$, and the corresponding systolic band Σ_γ of paths in $M^{\sim}$, one may define a (global) potential function u_γ axiomatically as follows.

Definition 3.1. — *Given a closed, orientable surface M of genus ≥ 2 with a Riemannian metric g, a non-trivial element $\gamma \in \pi_1(M, x_0)$ and the corresponding systolic band Σ_γ of complete, shortest-length paths in $M^{\sim}$ directed by γ, a global potential function $u_\gamma : M^{\sim} \to \mathbf{R}$ directed by γ is a function that satisfies the following axioms.*

(1) *For each $x^{\sim} \in M^{\sim}$, the function u_γ satisfies the relation*

$$u_\gamma(\gamma(x^{\sim})) = u_\gamma(x^{\sim}) + Sys_{\overline{\gamma}}(M, g) \ .$$

(2) *The function u_γ is of Lipschitz class everywhere in $M^\sim$, its Lipschitz constant satisfies $0 < Lip(u_\gamma, x^\sim) \leq 1$ at each $x^\sim \in M^\sim$; in the complement of the trace $B_{\gamma^\sim}$ of Σ_γ, the function u_γ is of class $C^{1,1}$ and its differential $du_\gamma(x^\sim)$ has norm ≤ 1 with respect to g, while if $x^\sim$ lies in the trace $B_{\gamma^\sim}$ of Σ_γ, then $Lip(u_\gamma, x^\sim) = 1$, and the directional derivatives of u_γ take the value 1 precisely in the direction of any path in Σ_γ that passes through $x^\sim$.*

(3) *For any given constant δ with $0 < \delta \leq 1$, there exists a positive constant C such that, for each point $x^\sim \in M^\sim$ at a distance $\geq C$ from the nearest path in Σ_γ the function u_γ satisfies $Lip(u_\gamma, x^\sim) \leq \delta$.*

(4) *The function u_γ has no critical points, in the sense that each level set of u_γ is a rectifiable, connected, properly imbedded curve in $M^\sim$, and for each $x^\sim \in M^\sim$, the image in M_γ under Π'_γ of each path of steepest ascent (respectively, descent) of u_γ from $x^\sim$ with $\gamma(x^\sim)$ (respectively, $\gamma^{-1}(x^\sim)$) is contained in a compact set, invariant under the translation by γ.*

Since addition of constants to potential functions does not affect the properties that characterize them as such, one may include an additional requirement that they vanish at a designated base point $x_0^\sim \in M^\sim$. To any potential function u_γ in $M^\sim$ one associates the corresponding reduced potential function $\overline{u}_\gamma : M_\gamma \to \mathbf{R}/(Sys_{\overline{\gamma}}(M,g)\mathbf{Z})$,

$$\overline{u}_\gamma(\Pi'_\gamma(x^\sim)) = u_\gamma(x^\sim) mod \cdot (Sys_{\overline{\gamma}}(M,g)\mathbf{Z}) \ . \tag{3.1}$$

The proof of the existence of a potential function u_γ for each non-trivial $\gamma \in \pi_1(M, x_0)$ when the metric has the required full generality is too long and technical to be fully included in the present paper ; however the initial step of a construction of these functions is easy and achieves the purpose, if the metric of (M,g) is smooth, at least of class $C^{1,1}$, and if the geodesics in $M^\sim$ have no conjugate points. This part of the proof is included for heuristic reasons.

Let (M,g) be an arbitrary closed, oriented surface of genus $g \geq 2$ with a metric g of class G ; let $M^\sim$, $g^\sim$, $x_0^\sim$, and γ be as before, and consider the systolic band Σ_γ of paths of shortest length in $M^\sim$ directed by γ. For any given path $\varphi^\sim \in \Sigma_\gamma$, choose an auxiliary base point $y_0^\sim \in \varphi^\sim$ as its initial point and parametrize $\varphi^\sim$ by its oriented arc length s from $y_0^\sim$; thus, for any real number s, we denote by $y^\sim(s)$ the point of $\varphi^\sim$ at an oriented distance s from $y_0^\sim$. Denoting by $d(x^\sim, y^\sim)$ the Riemannian distance between $x^\sim$ and $y^\sim$, the Busemann functions v_+ and v_- determined by the data of γ, $x_0^\sim$, $\varphi^\sim$, and $y_0^\sim$, are the real valued functions on $M^\sim$ defined by

$$v_+(x^\sim) = \lim_{s\to+\infty} (d(x^\sim, y^\sim(s)) - s) \ , \tag{3.2}$$

$$v_-(x^\sim) = \lim_{s\to-\infty} (d(x^\sim, y^\sim(s)) + s) \ .$$

By means of these functions, one introduces the functions u_γ and $h_{\gamma,\varphi^\sim}$, described respectively as the preliminary potential and stream functions, defined as follows:

$$u_\gamma(x^\sim) = \frac{1}{2}(v_-(x^\sim) - v_-(x_0^\sim) - v_+(x^\sim) + v_+(x_0^\sim)) \ , \tag{3.3}$$

$$h_{\gamma,\varphi^\sim}(x^\sim) = \frac{1}{2}(v_-(x^\sim) + v_+(x^\sim)) \ . \tag{3.4}$$

The following list of properties of the functions just introduced are either elementary, so that their proofs may be omitted.

1. The Busemann functions $v_\pm$ satisfy the Lipschitz condition $Lip(v_\pm) = 1$ and the functional identities

$$v_+(\gamma(x^\sim)) = v_+(x^\sim) - Sys_{\overline{\gamma}}(M, g) \ ,$$

$$v_-(\gamma(x^\sim)) = v_-(x^\sim) + Sys_{\overline{\gamma}}(M, g) \ ;$$

furthermore, if the metric is smooth and if geodesics have no conjugate points, the Busemann functions are of class C^1, with $|\bigtriangledown V_\pm| = 1$ and $\bigtriangledown V_- \neq -\bigtriangledown V_+$ everywhere.

2. For any fixed γ and under different choices of $x_0^\sim \in M^\sim$, $\varphi^\sim \in \Sigma_\gamma$, and $y_0^\sim \in \varphi^\sim$, the resulting functions v_+, v_-, u_γ, and $h_{\gamma,\varphi^\sim}$ are modified by additive constants.

3. The preliminary stream function $h_{\gamma,\varphi^\sim}$ is invariant under the action of γ ; it vanishes identically on the path $\varphi^\sim$ and takes values ≥ 0 everywhere else in $M^\sim$; furthermore it satisfies everywhere the Lipschitz condition $Lip(h_{\gamma,\varphi^\sim}) \leq 1$. If the

metric is smooth and if there are no conjugate points, then $\bigtriangledown h_{\gamma,\varphi^\sim}$ exists everywhere, is continuous, and $| \bigtriangledown h_{\gamma,\varphi^\sim}| < 1$.

4. The preliminary potential function $u\gamma$ satisfies the Lipschitz condition $Lip(u_\gamma) \leq 1$ everywhere in $M^\sim$; in addition it satisfies the relation $u_\gamma(\gamma(x^\sim)) = u_\gamma(x^\sim) + Sys_{\overline{\gamma}}(M,g)$; it is normalized additively so that it vanishes at $x_0^\sim$, and depends only on γ, not on the particular choice of path $\varphi^\sim \in \Sigma_\gamma$. If the metric is smooth and if there are no conjugate points, then $\bigtriangledown u_\gamma$ exists everywhere, is continuous, and is nowhere zero.

If the metric g is sufficiently smooth and if the geodesics in $M^\sim$ have no conjugate points, then the gradient flow of the function u_γ constitutes a foliation of $M^\sim$, including each of the paths $\varphi^\sim \in \Sigma_\gamma$. The dynamical system of this flow is invariant under the translation group generated by γ and the preliminary stream function $h_{\gamma,\varphi}$ is constant along each orbit. In addition, since $\bigtriangledown h_{\gamma,\varphi^\sim} = 0$ only along complete geodesics of the systolic band Σ_γ, it follows that, under the special assumptions on g, the preliminary stream functions $h_{\gamma,\varphi^\sim}$ are actual stream functions, constant along, and locally separating, the orbits of the gradient flow of u_γ, so that each of these orbits is invariant under the translation of γ. This shows that, if the metric g is of class at least $C^{1,1}$ and if there are no conjugate points, the function u_γ satisfies all the four properties characterizing a global potential function directed by γ. In the general case, the potential functions can be obtained similarly from formulas (3.3), (3.4), in which the Busemann functions $V_\pm$ are replaced by corresponding functions with similar properties, but constructed by a process yielding functions better suited to our purposes. An important fact, applied in that construction is the uniform exponential growth property of $\pi_1(M, x_0)$ in terms of its generators.

Given a non-trivial $\gamma \in \pi_1(M, x_0)$, consider the covering surface M_γ of M the cyclic subgroup of $\pi_1(M, x_0)$ generated by γ with the metric induced by the covering map from g. The surface M_γ is homeomorphic to a cylinder, with $M^\sim$ as its universal covering surface. The family Σ_γ of paths of least length in $M^\sim$ directed by γ corresponds to the family of systole-long paths in M_γ, also denoted by Σ_γ. Since the genus g of M is assumed to be ≥ 2, the family Σ_γ is compact ; if it consists of more than one essential closed path in M_γ, there are two such paths, bounding a retract of M_γ, that contains all the systole-long paths in M_γ. This retract will be called the *systolic strip* of γ, denoted by Y_γ ; if γ is a "simple" element of $\pi_1(M, x_0)$, meaning that γ is representable by a simple loop in M, then the covering map Π_γ of M by M_γ is a one-to-one isometry of Y_γ onto its image in M ; the latter will likewise be denoted by Y_γ, despite the small risk of ambiguity. This is the case, in particular, if $\overline{\gamma}$ is a critical, free homotopy class for the systole of (M, g). We observe that, for general metrics g, the systolic strips Y_γ, whether considered in M_γ or in M, are not necessarily covered by the paths of the systolic band Σ_γ.

4. THE PRELIMINARY VARIATIONAL PROCESS

We shall consider now the problem of characterizing extremal isosystolic metrics in a closed, orientable surface M of genus $g \geq 2$ in terms of local properties such as, for instance, solutions of partial differential equations. The surface M is assumed to be *polarized*, meaning that the fundamental group $\pi_1(M, x_0)$, regarded as the group of homotopy classes of x_0-based loops, is identified with its standard presentation as an abstract group, by an explicit choice of $2g$ generators $p_i, q_i (1 \leq i \leq g)$, satisfying the relation

$$p_1 q_1 p_1^{-1} q_1^{-1} \cdots p_g q_g p_g^{-1} q_g^{-1} = e \ , \tag{4.1}$$

in the usual way, first by orienting M, then by assigning to each of the $2g$ abstract generators p_i, q_i a corresponding system of $2g$ x_0-based, simple, oriented loops, that are pairwise disjunct away from x_0, chosen in such a way that, if one cuts M along these $2g$ loops, one obtains a simply connected domain D. In addition, the loops are constructed, so that one may read the oriented boundary of D as a sequence of $4g$ oriented loops in M corresponding to the left-hand side of the relation (4.1).

The first question that arises in considering the extremal isosystolic problem is that of characterizing the subsets $S = S_g \subset \pi_1(M, x_0)$ that may possibly occur as the *critical* subsets for some extremal isosystolic metric g: the sets S_g consist of the elements $\gamma \in \pi_1(M, x_0)$, whose conjugacy classes $\overline{\gamma}$ are represented by its systole-long paths. The question is somewhat ambiguous as posed, because extremal isosystolic metrics are not necessarily continuous in M. In fact, consider a surface M with an extremal isosystolic metric g, normalized under the condition $Sys(M, g) = 1$: there may be (cf. the second example in Section 6) a sufficiently small constant $\delta > 0$ and a non-empty family of non-critical homotopy classes γ, whose corresponding local systoles $Sys_{\overline{\gamma}}(M, g)$ lie in the interval $1 < Sys_{\overline{\gamma}}(M, g) \leq 1 + \delta$, and are achieved by closed paths $\varphi_{\overline{\gamma}}$. Choosing one (or a finite set) of such paths $\varphi_{\overline{\gamma}}$, one may replace the metric g with a discontinuous metric g', identical with g outside the trace of the chosen path(s) $\varphi_{\overline{\gamma}}$, and uniformly smaller than g along each $\varphi_{\overline{\gamma}}$, that reduces the length of $\varphi_{\overline{\gamma}}$ to unity, thereby rendering it of systole length: if δ is sufficiently small, the change of metric will not affect the length-minimizing property of the paths that are systole-long in terms of g. The metric g' then is again an extremal isosystolic one, having the same systole and total area as g, but its critical set S', interpreted literally, would include S together with all the homotopy classes γ affected by the change. The following definition is proposed in order to clarify the ambiguity.

Definition 4.1. — *Let (M, g) be a closed surface of genus $g \geq 2$ with a Riemannian metric g that is extremal isosystolic. Then the* <u>*essential critical set*</u> *with respect to g is the subset $S = S_g \subset \pi_1(M, x_0)$ consisting of the homotopy classes γ, such that the conjugacy classes $\overline{\gamma}$ of each $\gamma \in S$ are represented each by a band of systole-long paths, whose trace in M has* <u>*positive measure*</u>*. A* <u>*maximal critical set*</u> *with respect to g is a set $S' = S'_g$, $S_g \subseteq S'_g \subset \pi_1(M, x_0)$, that is maximal with respect to the property of being representable by systole-long paths in terms of any metric g' such that $g' = g$ almost everywhere and each systolic band of g is also a systolic band of g', so that the values of both the area of M with respect to either metric coincide.*

One observes that any extremal isosystolic metric g in M may be represented as the limit, almost everywhere, of an increasing sequence (g_n) of smooth Riemannian metrics, such that the isosystolic critical set of free homotopy classes of each g_n is, for instance, a maximal critical set (or else, trivially, an essential one) of classes with respect to g. The following proposition describes some general properties of the subsets $S \subset \pi_1(M, x_0)$ that may occur as the set of critical classes (essential or maximal) with respect to an extremal isosystolic metric.

Proposition 4.2. — *If g is an extremal isosystolic metric in M, then the following assertions are valid.*

(i) *The essential critical subset $S = S_g \subset \pi_1(M, x_0)$ with respect to g is necessarily the union of a finite family of non-trivial conjugacy classes, symmetric with respect to inversion.*

(ii) *Each element of S is representable by a simple, closed path.*

(iii) *Any two elements of S may be represented by two simple, closed paths of least length, that either cross each (transversally) at exactly one point, or are disjunct, or may be approximated uniformly by two disjunct, simple, closed paths.*

(iv) *For each element of S there is at least one other element of S, such that any pair of simple, closed curves representing the two respective classes have a non-empty intersection.*

(v) *(conjectured) The canonical image of S in the homology group $H_1(M, \mathbf{Z})$ includes a family of $2g$ elements constituting an integral basis of $H_1(M, \mathbf{Z})$.*

Proof. The finiteness and symmetry properties (i) of S are trivial. The intersection properties (ii) and (iii) are well known, elementary properties, that are verifiable by suitable, smooth approximations of length minimizing paths in surfaces with path-length metrics (and by the geodesic paths themselves, if the metric is smooth). The symplectic property (iv) is a consequence of Lemma 2.1 and Property (iii).

Property (v), the conjectured property of homology fullness of the set S, is the only one that requires some comment. Even though it may seem intuitively obvious, in actual fact it is essentially equivalent to the strict monotonicity of the minimum isosystolic ratio as a function of the genus of the surface M. Current attempts to prove the strict monotonicity of the minimum isosystolic ratio depend on sharp numerical estimates of the "systole relative to the boundary" for surfaces with boundary, with an extremal isosystolic metric g_c $(0 < c < 3)$, in which the homology systole is unity, and c is the minimum length of any curve homotopic to the boundary. A numerical computation of these estimates is now in progress (cf. the first example in Section 6).

A consequence of Gromov's compactness theorem [5] is that, for any given value $g \geq 2$ of the genus of M, there can be at most finitely many subsets $S \subset \pi_1(M, x_0)$, pairwise inequivalent under the action of the automorphism group of $\pi_1(M, x_0)$, such that each is the set of critical homotopy classes for some extremal isosystolic metric in M. On the other hand the properties on the sets S stated in Lemma 4.1 do not appear to be even sufficient to deduce that there are finitely many sets S, up to equivalence, that verify these properties. The problem of characterizing *a priori* the sets of critical homotopy classes, either essential or maximal, of extremal isosystolic metrics for surfaces of genus ≥ 2 is open and probably very difficult. For the present purposes it is sufficient to remark that, given any set $S_0 \subset \pi_1(M, x_0)$ with the five properties listed in Proposition 4.2, there are at most finitely many sets S' with $S_0 \subseteq S' \subset \pi_1(M, x_0)$, verifying the same properties.

The "variational search" for an extremal isosystolic metric in M consists of two "processes", performed alternately infinitely many times, where each process replaces a given, non-extremal metric with another metric, exhibiting a lower isosystolic ratio. The present section outlines the first process, while Section 5 is devoted to describing the second one.

The first process consists by itself of an infinite iteration of two alternating steps, Steps 1 and 2, described below.

Step 1. Given any admissible metric g_0 in M (a smooth one, to begin with), normalized by the condition $Sys(M, g_0) = 1$, one introduces the infinite dimensional numerical torus $\mathbf{T} = \Pi_{\gamma \in \pi_1(M,x_0)\setminus\{e\}}(\mathbf{R}/\mathbf{Z})$, whose components are indexed by the non-trivial elements of the fundamental group of $\pi_1(M, x_0)$. One then considers the mapping $F : M^\sim \to \mathbf{T}$, where $M^\sim$ is the universal covering surface of M, defined by

$$(4.2) \qquad y_\gamma(x^\sim) = \{(Sys_{\overline{\gamma}}(M, g_0))^{-1} u_\gamma(\Pi'_\gamma(x^\sim))\} mod \mathbf{Z} \ \ ; \ F = (y_\gamma)_{\gamma \in \pi_1(M,x_0)\setminus\{e\}} \ ,$$

In this equation, for each $\gamma \in \pi_1(M, x_0)\setminus\{e\}$, the function $u_\gamma : M_\gamma \to \mathbf{R}$ is a potential function directed by γ, as in definition (3.1). Clearly each component function y_γ of F is of Lipschitz class, with a Lipschitz ratio $\leq (Sys_{\overline{\gamma}}(M, g_0))^{-1} \leq 1$ at each point. The mapping F is equivariant under the diagonal action of $\pi_1(M, x_0)$ on $M^\sim \times \mathbf{T}$ as a transformation group, acting simultaneously by translation on $M^\sim$, and on $\mathbf{T}$

by permuting the components according to the adjoint action of the group on itself. Therefore (see the lemma that follows) there exists a unique Riemannian metric $g_1^{\sim}$ on $M^{\sim}$, determined at each point $x^{\sim}$ by the property of having the least possible element of volume, subject to the constraint that, for each γ, the weak differential [3] $dy_\gamma(x^{\sim})$ has a norm ≤ 1 with respect to $g_1^{\sim}$; since this metric is obviously invariant under translations of $M^{\sim}$ by $\pi_1(M, x_0)$, it induces a metric g_1 in M with stystole ≥ 1 and area no larger than the original one with respect to g_0. It is precisely this step that would be carried out more naturally in the context of Finsler metrics, while the restriction to Riemannian metrics introduces some complications. The preference given here to Riemannian metrics is due to the greater familiarity with Riemannian geometry by most people.

The result of the change of metric just described may alter, in general, the set of isosystolic critical classes, by either addition or deletion ; furthermore the potential functions with respect to g_0 for each $\gamma \in \pi_1(M, x_0)\backslash\{e\}$ need not satisfy the axioms set in Definition 3.1 in terms of g_1: in particular the condition $|dy_\gamma(x^{\sim})|_{g_1} = 1$ is no longer equivalent, in general, to the property of $x^{\sim}$ lying in the trace of the systolic band directed by an essential critical class γ.

Step 2. The second step of the process consists of replacing the family of potential functions u_γ with respect to of the original metric g_0 with a corresponding family in terms of g_1, thereby altering the mapping $F : M^{\sim} \to \mathbf{T}$.

One then repeats this two-step operation. It follows once more from the compactness theorem for the isosystolic problem, that, after a finite number of iterations, the operation first stabilizes the set of critical classes of paths, and leads to a sequence of metrics, converging to a metric g, such that neither of the two steps just outlined necessarily lead to any further change. Such a metric g is characterized by the three following properties:

[3] A Lipschitz-continuous function f is said to be non-critical at an interior point $x{\in}M$, if every neighborhood of x contains an open neighborhood U of x, within which both subsets $\{y{:}f(y){>}f(x)\}$ and $\{y{:}f(y){<}f(x)\}$ are non-empty and contractible. The differentials of two such functions f and g are said to be weakly linearly independent at x, if, for each pair of constants $(a{,}b){\neq}(0{,}0)$, the function $af{+}bg$ is non-critical at x. The weak differential $df(x)$ of f at x is the closure of the set of all differentials at x of smooth functions g, such that $f{-}g$ is critical at x. It follows that the set $df(x)$ is non-empty and is tightly contained in the euclidean ball of radius $Lip(f{,}x)$.

(i) the metric g is uniquely determined by the mapping $F : M^{\sim} \to \mathbf{T}$ defined in (4.2), under the property of minimizing almost everywhere the element of area, subject to the constraint $|du_\gamma(\Pi'_\gamma(x^{\sim}))|_g \leq 1$ for every $x^{\sim} \in M^{\sim}$ and every γ ;

(ii) for any given $x^{\sim}$ and γ, $|du_\gamma(\Pi'_\gamma(x^{\sim}))|_g = 1$, if, and only if $x^{\sim}$ lies in the systolic band in $M^{\sim}$ directed by γ ;

(iii) The systolic strips cover M in such a way that almost every point $x \in M$ is a transversal intersection point of at least two systole-long paths, and the angles between the successive, oriented, systole-long paths meeting at x are each $\leq \pi/2$ intersection, with a total angle $\geq 2\pi$ (the proof essentially the same as in Lemmas 2.1, 2.2 and 2.4).

Recall now the definition of the systolic strips, given in the closing paragraph of Section 3: for each non-trivial homotopy class $\gamma \in \pi_1(M, x_0)$ the systolic strip Y_γ directed by γ is a compact retract of M_γ (the covering space of M corresponding to the cyclic subgroup $\langle\gamma\rangle \subset \pi_1(M, x_0)$), obtained by "filling in" the compact subdomain between the two outermost systole-long paths in M_γ. Restricting γ now to the critical isosystolic classes of M, the corresponding systolic strips may be identified with the corresponding images in M under the covering map, one sees that M is covered (more descriptively, "bandaged") by this finite family of critical systolic strips, so that each point is covered by at least one pair of mutually transversal strips, and every tangent vector makes an angle $\leq \pi/4$ with a gradient vector of the potential function at least one critical class. The total surface M is thus decomposed into a finite number of compact, convex, geodesic polygons, partially ordered by inclusion, each one determined by some interior point x, and defined as the component of x in the intersection of all the critical systolic strips containing x in their interior. Each of these geodesic polygons, determined by the intersection of k unoriented systolic strips (or $2k$ oriented ones), is completely described by the corresponding k pairs of mutually opposite potential functions, and is handled in the next section in a way that generalizes the treatment of k-regular domains in Section 2.

5. THE EULER-LAGRANGE EQUATIONS

Let M be a closed, orientable surface with a geodesic Riemannian metric g, such that the critical systolic strips cover M with the conditions described at the end of Section 4, and let D be a geodesic polygon in M, determined by a finite intersection of k unoriented, critical systolic strips, and such that there exists some interior point of D that is not interior to any additional, critical systolic strip. Let $\pm u_1, \pm u_2, \cdots, \pm u_k$ denote the corresponding potential functions and

$$\mathbf{u} = (u_1, u_2, \cdots, u_k) : D \to \mathbf{R}^k$$

the mapping that they define, where the choice of orientation of each component u_j is immaterial. The Riemannian metric in D, by assumption, is the one which, at almost every point $x \in D$, has the least area density form under the condition $|di_j(x)|_g \leq 1$ for each j $(1 \leq j \leq k)$; in particular the metric at x is controlled by the subset among the functions u_j such that $|du_j(x)|_g = 1$. If two of these functions (let us say u_1 and u_2) suffice to determine g in an open subset $U \subset D$, then du_1 and du_2 are mutually orthogonal (by Lemma 2.3) ; otherwise, for almost all x there is a neighborhood U of x, where three of them, say u_1, u_2 and u_3 determine g by themselves under the condition $|du_j(x)|_g = 1$ $(j = 1, 2, 3)$, and the six disjunct angles in terms of g formed by the six differentials $\pm du_j$ in the cotangent space of x are all $< \pi/2$ (cf. Lemma 2.4). In any case the mapping $\mathbf{u}$ of equation (5.1) is almost everywhere an immersion of D, inducing a metric, that is determined at each point by the direction of the tangent plane. This situation is precisely one considered by É. Cartan in his 1933 monograph [5] for the purpose of studying invariants attached to variational problems. What makes the isosystolic problem difficult from the viewpoint of Cartan's treatment is that the area functional is not of class C^2 in terms of the direction parameters of the tangent plane. For instance, if $k = 3$, suppose that the image $\mathbf{u}(D)$ in $\mathbb{R}^3$ is represented locally as the graph of a function, $u_3 = f(u_1, u_2)$, or, more briefly, $z = f(x, y)$, and denote by (u, v, w) the homogeneous direction numbers of any plane at the point (x, y, z) ; then the tangent plane of $\mathbf{u}(D)$ at (x, y, z) has direction numbers proportional to $(p, q, -1)$, where $p = \partial f/\partial x$ and $q = \partial f/\partial y$). It follows from the assumptions $|dx|_g = |dy|_g = 1$ that the area functional $dVol_g$ is expressed by (2.1), (2.2). In order to account for the alternative possibilities that the least-area metric be controlled by all three of the potential functions or just two of them, it is best to express the area density function in terms of the three homogeneous direction numbers $\xi = (\xi_1, \xi_2, \xi_3)$ of the tangent plane, by means of a function $L(\xi)$, defined to be positively homogeneous of degree one, such that $dVol_g = L(p, q, -1)$. Introduce the function (Heron's function)

$$\Delta(\xi_1, \xi_2, \xi_3) = Re\{(2(\xi_2^2\xi_3^2 + \xi_3^2\xi_1^2 + \xi_1^2\xi_2^2) - (\xi_1^4 + \xi_2^4 + \xi_3^4))^{\frac{1}{2}}\} ,$$

representing, when it is positive valued, twice the area of the euclidean triangle with sides of length $|\xi_1|$, $|\xi_2|$, $|\xi_3|$; then direct computation leads to the following expression for $L(\xi)$:

$$L(\xi) = \begin{cases} 2|\xi_1\xi_2\xi_3|/\Delta(\xi_1, \xi_2, \xi_3) \ , & \text{if } \mathrm{Max}\{\xi_1^2 \ , \xi_2^2, \xi_3^2\} < \frac{1}{2}(\xi_1^2 + \xi_2^2 + \xi_3^2) \\ \mathrm{Max}\{|\xi_1|, |\xi_2|, |\xi_3|\}, & \text{if } \mathrm{Max}\{\xi_1^2, \xi_2^2, \xi_3^2 \geq \frac{1}{2}(\xi_1^2 + \xi_2^2 + \xi_3^2) \end{cases}$$

The two alternative analytic expressions given above for $L(\xi)$ reflect the two respective possibilities, whether the metric is controlled by all three of the potential functions or by just two of them. The global function $L(\xi)$ represents geometrically the least diameter of any disk that contains a triangle with sides of length $|\xi_1|$, $|\xi_2|$, $|\xi_3|$, and is the lower convex envelope of the one expressed by the algebraic function in the top line.

Consider now a small variation of the immersion $\mathbf{u}$ in a compact subdomain of D, reflected by a corresponding variation of the functional relation between x, y, z in a compact subdomain of its domain of definition. Such a change induces a corresponding change in the metric in the global surface M, that generally leaves all critical local systoles unchanged, and, if the change is sufficiently small, does not introduce any new critical homotopy classes. Thus the isosystolic variational problem is reduced, locally, to that of finding functions $z = f(x, y)$ with prescribed boundary values in a compact domain $D \subset \mathbb{R}^2$, that minimize the integral $\int_D L((\partial f/\partial x, \partial f/\partial y, -1)|dx \wedge dy|$. A routine calculation yields readily the Euler-Lagrange equation for this problem, in non-parametric form:

$$\frac{\partial^2 L(p,q,-1)}{\partial p^2}\,\frac{\partial^2 f(x,y)}{\partial x^2} + 2\,\frac{\partial^2 L}{\partial p \partial q}\,\frac{\partial^2 f(x,y)}{\partial x \partial y} + \frac{\partial^2 L}{\partial q^2}\,\frac{\partial^2 f(x,y)}{\partial y^2} = 0 \ . \tag{5.3}$$

The convexity of the homogeneous function $L(\xi)$ guarantees that the local boundary value problem has a unique, weak solution of class $C^{1,1}$, that assumes the assigned boundary values, provided that the domain is strictly convex. On the other hand $L(\xi)$, in contrast with the classical analogue $L_0(\xi) = ((L\xi_1^2 + \xi_2^2 + \xi_3^2))^{\frac{1}{2}}$ arising from Plateau's problem, is not strictly convex in every non-homogeneous plane in $\mathbb{R}^3$. Indeed it is strictly convex, locally, only in non-homogeneous planes at points where the top-line formula of (5.2) holds, i.e. in the open domain $\Omega \subset \mathbb{R}^3$, where ξ_1^2, ξ_2^2 and ξ_3^2 satisfy the triangle inequality. For any solution $f(x, y)$ of the Euler-Lagrange equation (5.3), the solution is of elliptic type, and hence real analytic, at any point of D for which $\xi \in \Omega$.

As the point for which $\xi \in \Omega$ approaches a point (x, y, z) for which ξ lies on a regular part of the boundary of Ω (consisting of all lines through the origin, making an angle of 45° with exactly one of the three coordinate axes), the solution of (5.3) becomes a degenerate elliptic solution, approaching a parabolic point, where the bicharacteristics are tangent to the curve of parabolic points, Hence, if the analytic solution of the elliptic equation admits an analytic continuation across such a transition curve, it provides an example of solution of a non-linear differential equation of Tricomi type. However, the coefficients of the actual equation (5.3) and the discontinuity of the second derivatives of L at the boundary of Ω imply that, when ξ lies in the interior of the complement of Ω (consisting of six disjunct, convex, open, quadric cones), then the function $f(x, y)$ is not restricted by the Euler-Lagrange equation at all ; indeed two of the three potential functions constitute a local cartesian chart for a flat metric g, and the third one is only required to have a Lipschitz constant ≤ 1 in terms of g. We summarize the main results of preceding discussion in the statement of the following theorem.

Theorem 5.1. — *Let (M, g) be a closed surface with an extremal isosystolic metric g, and let $\overline{\gamma}_j^{\pm 1}$ $(j = 1, 2, 3)$ be six distinct, critical free homotopy classes of closed paths in M, admitting an open, simply connected region $D \subset M$ that does not meet any critical isosystolic strip other than the six strips directed by $\overline{\gamma}_j^{\pm 1}$. Let $u_j : D \to \mathbb{R}$ be local potential functions in D directed by each $\overline{\gamma}_j$: then the combined map (u_1, u_2, u_3): $D \to \mathbb{R}^3$ is an immersion, that may be described locally in non-parametric form as* *the graph of a function* $z = f(x, y)$ *as* (x, y, z) *traces the image* $(u_1, u_2, u_3)(D) \subset \mathbb{R}^3$. *In addition the function* f *satisfies the Euler-Lagrange equation (5.3). The latter equation is vacuous in the interior of any subdomain of* D*, whose image in* $\mathbb{R}^3$ *has the property that each of its tangent planes makes an euclidean angle* $< 45°$ *with any one of the three coordinate planes. In the interior of any region for which the tangent planes make angles* $> 45°$ *with each of the three coordinate planes, the equation (5.3) is elliptic.*

One distinguishes, among the elliptic solutions of the Euler-Lagrange equation (5.3), the trivial class, consisting of potential functions that are related by linear equations: the metric that they induce is obviously flat. Conversely, it is not hard to verify that, if an extremal isosystolic metric, determined by three potential functions related by an elliptic solution of (5.3), is locally flat, then the functions must be linearly related.

In order to complete the variational analysis, one must still consider the domains of M, where an extremal isosystolic metric may be controlled by four or more overlapping, unoriented isosystolic bands. Clearly, the functional relation between any three local potential functions (directed by pairwise transversal critical isosystolic bands) may be expressed locally by choosing two of them as local parameters, so that the third one is represented as a function of two variables, namely $z = f(x, y)$, in the discussion leading to Theorem 5.1. If $f(x, y)$ satisfies the Euler-Lagrange equation (5.3) in a domain D, then the local potential functions directed by any number of additional critical isosystolic classes could conceivably enter the local picture without disturbing the equilibrium condition implied by the Euler-Lagrange equation: in fact, given any smooth Riemannian metric in D, any smooth curve in D is locally an orthogonal trajectory of a uniquely determined, local geodesic foliation, that may generate such a potential function. One can prove that it is not possible to have two functions $f_1(x, y)$ and $f_2(x, y)$ that are simultaneously elliptic solutions of (5.3), inducing the same volume form (and hence the same metric), unless f_1 and f_2 are being either both linear functions of x and y, or else differ from each other by an additive constant, or have a constant value for their sum. It seems plausible to conjecture that the interior extremality condition in the domain D of intersection of four or more isosystolic strips should lead to a decomposition of D into compact sub-regions, in each of which only three of the potential functions would be related by (5.3). In addition, it still seems very hard to investigate the regularity (or non-regularity) of the interface curves between the resulting triple intersections of systolic bands of strips. The next section is the result of an attempt to determine the free boundary that arises between subdomains where a solution of (5.3) is of elliptic type on one side, while, on the opposite side, the extremal metric is controlled by just two potential functions.

6. A SPECIAL FREE BOUNDARY PROBLEM

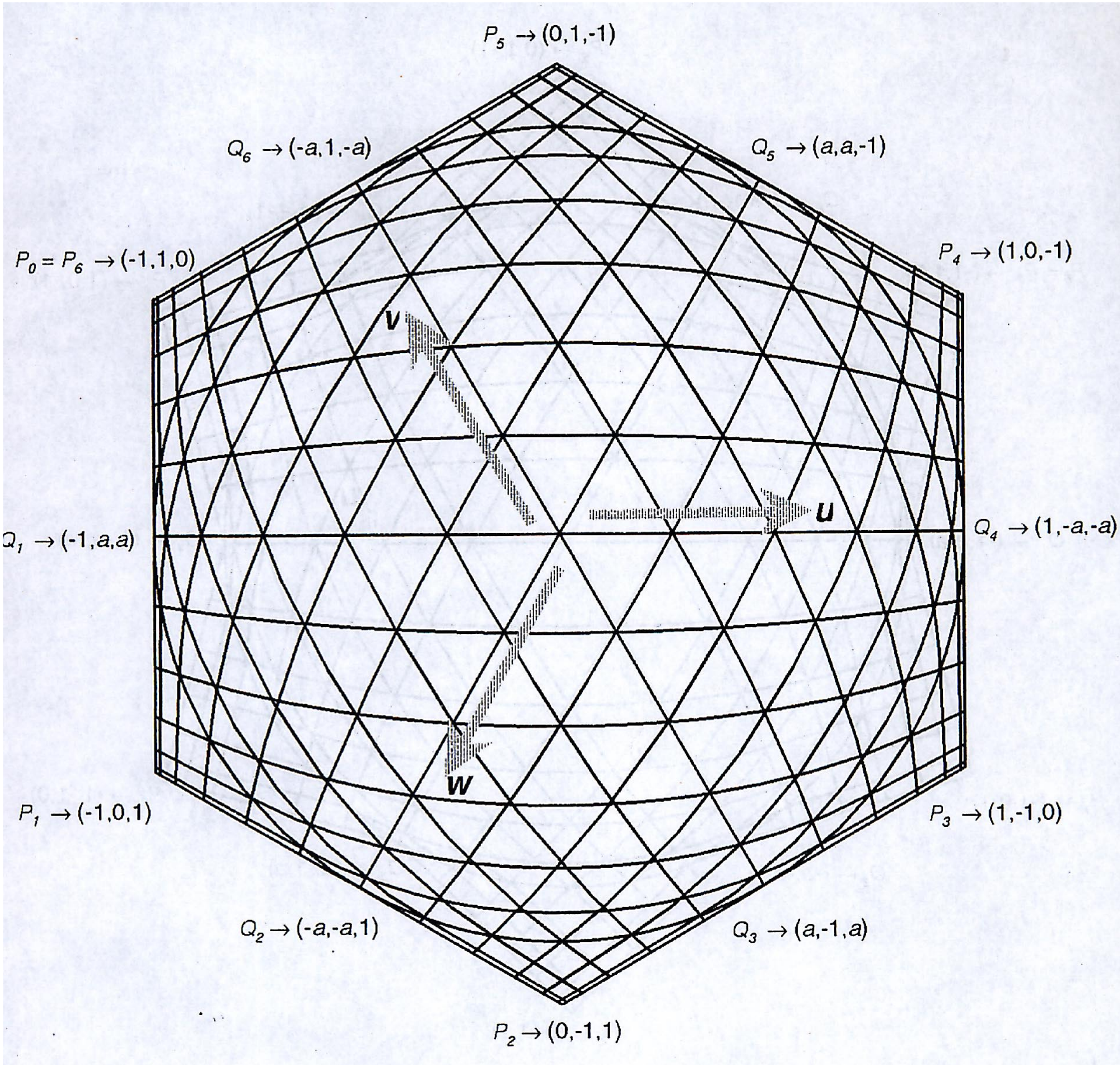
The following boundary value problem related to the extremal isosystolic metric problem leads to a special, non-elementary, local solution of the Euler-Lagrange equation (5.3) ; there is some hope that this solution may be evaluated explicitly (in the sense of classical function theory), thus providing a useful comparison barrier for estimates of more general solutions.

Problem. — Given a closed, regular euclidean hexagon Ω, find a Riemannian metric in Ω that minimizes its total area, subject to the constraint that each point of each side of Ω has least Riemannian distance equal to 2 from the diametrically opposite side.

We postulate the following two additional properties of the metric sought in this problem, with the hope that at least one of them may be deduced as consequences of the other data.

(i) Opposite pairs of sides of the hexagon Ω have a one-to-one correspondence, so that each point of any one side of Ω is the only one that has distance $= 2$ from the corresponding point on the opposite side ;

(ii) the metric achieving the minimum area in the problem is unique, and hence admits the dihedral group D_6 of symmetries of the hexagon as its own group of isometries.



P5 → (0,1,-1)
Q6 → (-a,1,-a)
Q5 → (a,a,-1)
P0 = P6 → (-1,1,0)
P4 → (1,0,-1)
v
u
Q1 → (-1,a,a)
Q4 → (1,-a,-a)
w
P1 → (-1,0,1)
P3 → (1,-1,0)
Q2 → (-a,-a,1)
Q3 → (a,-1,a)
P2 → (0,-1,1)

Figure 1.

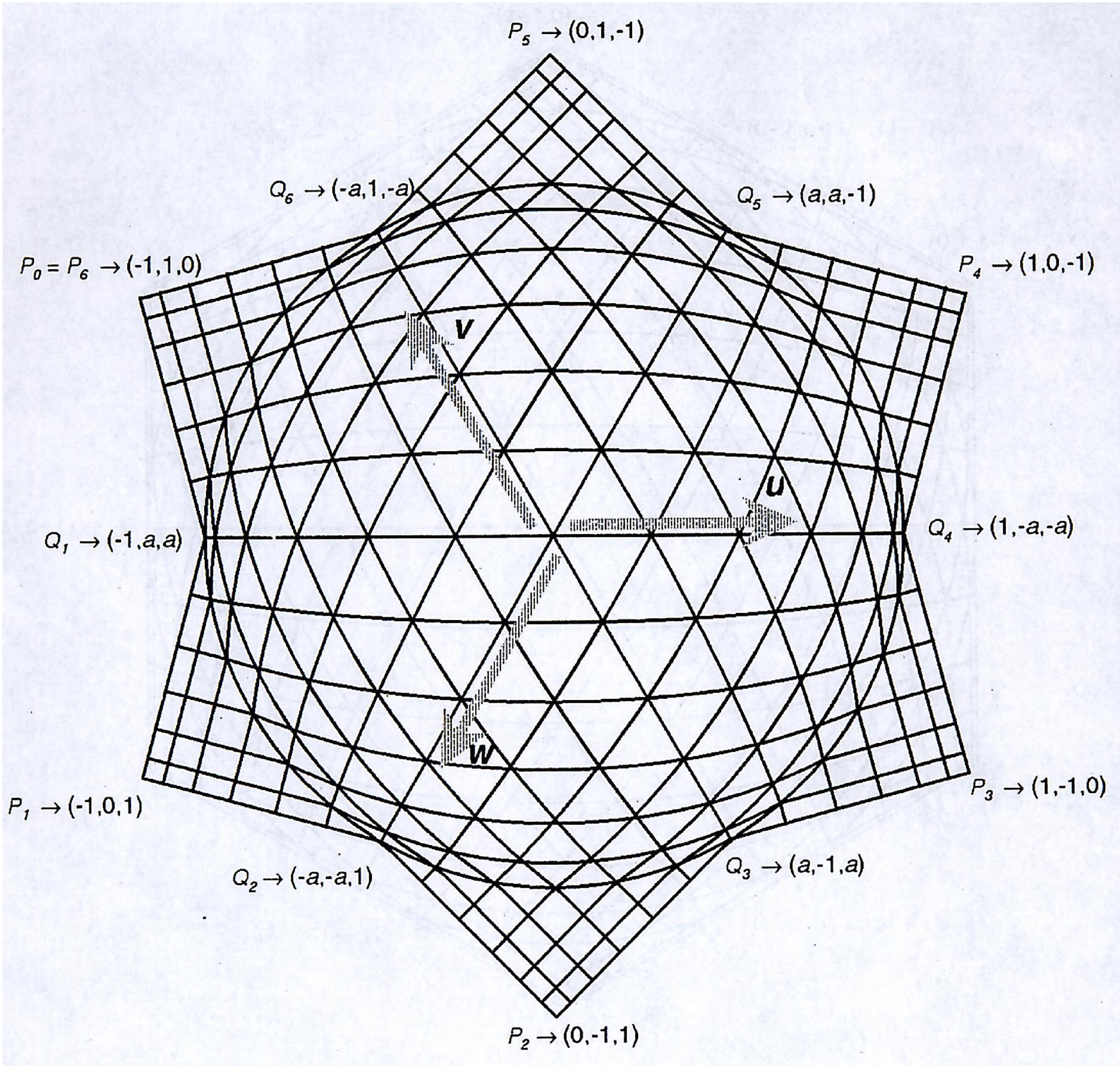

$P_5 \to (0,1,-1)$
$Q_6 \to (-a,1,-a)$
$Q_5 \to (a,a,-1)$
$P_0 = P_6 \to (-1,1,0)$
$P_4 \to (1,0,-1)$
V
U
$Q_1 \to (-1,a,a)$
$Q_4 \to (1,-a,-a)$
W
$P_1 \to (-1,0,1)$
$P_3 \to (1,-1,0)$
$Q_2 \to (-a,-a,1)$
$Q_3 \to (a,-1,a)$
$P_2 \to (0,-1,1)$

Figure 1a.

To begin with, we label and orient the six sides of Ω (see Figure 1) by orienting and tracing its boundary, starting and ending at a vertex $P_0 = P_6$, and denote each successive, oriented side by l_j ($j = 1, \cdots, 6$ *modulo* 6), in the order in which it is traced, so that l_j runs from P_{j-1} to P_j. According to the results of Section 4, the problem is equivalent to the following one. Consider any family $(u_j)_{j\in\mathbb{Z}/6\mathbb{Z}}$ of six real valued functions of Lipschitz class in Ω, such that $u_{j+3} = -u_j$, and satisfying the following four conditions.

(i) Each u_j ranges in the interval $[-1, 1]$, taking the values -1 and 1 respectively and exclusively on the sides l_j and l_{j+3}.

(ii) The functions u_1, u_3 and u_5, also renamed u, v and w, define an imbedding of class $C^{1,1}$

$$(u, v, w) : \Omega \to K^3 = \{(x, y, z) \in \mathbb{R}^3 : max\{|x|, |y|, |z|\} \leq 1\}$$

into the cube K^3 of edge length $= 2$.

(iii) The image surface is transversal to the direction of all three coordinate axes at each of its points.

(iv) The image surface $\Sigma = (u, v, w)(\Omega)$ is invariant under the group (D_6D_3) of rotations and reflections of $\mathbb{R}^3$ that map the cube K^3 and the line $\{x = y = z\}$ onto themselves.

Under these conditions, in particular, the image surface is the graph of a function $w = f(u, v)$ in a certain (as yet undetermined) subdomain $\Theta \subset \{(u, v) \in \mathbb{R}^2 : |u| \leq 1, |v| \leq 1\}$, with both first order partial derivatives $p = \frac{\partial f}{\partial u}$, $q = \frac{\partial f}{\partial v}$ strictly negative everywhere. The main problem is to find a function $f(u, v)$ that satisfies the partial differential equation (5.3), with $L(p, q - 1)$ defined in (5.2), such that its graph in K^3 satisfies the symmetry and the boundary conditions described above.

In the next few paragraphs we refer directly to Figure 1 and 1a ; the latter is a schematic representation of the hexagon Ω, indicating next to each labeled point the local values taken by the triple values taken by the triple (u, v, w) of potential functions. The thick, shaded arrows indicate the gradient flow of the indicated functions u, v and w near the center of the respective systolic bands, while the thin lines suggest the web of shortest paths connecting opposite sides.

Consider first the possible configuration of the six shortest paths $(\varphi_j)_{j\in\mathbb{Z}/6\mathbb{Z}}$ in Ω where φ_j is the shortest path from the vertex P_{j-1} to the vertex P_{j+1}. Each of these six paths constitutes one of the two boundary paths of the corresponding "systolic band" of shortest paths from the side l_{j-1} to the side l_{j+2}, and is also an orbit of the gradient flow of u_{j-1}. One sees that the union of these six "boundary" shortest paths must cover the whole boundary of the hexagon Ω, because otherwise there would be open sets in Ω, meeting at most one systolic band, contradicting Lemma 2.1. Recalling the postulated hexagonal symmetry of the solution of the problem, it is easily seen that, for each $j mod.6$, an initial segment of the path φ_j must trace at least the initial "half" of the side l_j (see Fig.1, i.e. the segment of l_j from P_{j-1} to the midpoint Q_j. By symmetry, a terminal segment of the path φ_j traces the terminal half of l_{j+1}, from Q_{j+1} to P_{j+1}.

On the other hand the paths φ_j cannot lie totally in the boundary of Ω . In fact, suppose that one of the paths φ_j, and hence all six, were to be entirely contained in $\partial\Omega$; in other words, suppose that each φ_j were to just trace the two sides l_j and l_{j+1}: this would mean that since each φ_j has length 2 and describes a shortest path between any two of its points, and because of the symmetry of Ω, the function $u_1 + u_3 + u_5 = u + v + w$ would be identically zero on the whole boundary $\partial\Omega$. Note that $u+v+w$ satisfies the same linearized form of the second order equation (4.8) as w, in terms of the four variables u, v, p and q and the equation would have to be of elliptic type everywhere in the interior of Ω ; hence $u+v+w$ satisfies the conditions for the maximum-minimum principle and, since it vanishes identically of the boundary of Ω, it follows that $u+v+w = 0$ identically in Ω. Therefore the gradient flows of the functions u_j would all be straight lines, contrary to the assumption. This argument shows that each vertex P_j has a neighborhood U_j in Ω, bounded by a segment ψ_j in the middle of the path φ_j, such that U_j is 2-regular (in the sense of Section 2). By the same token, the interior Ω_0 of $\Omega\backslash(\cup_{1\leq j\leq 6} w(U_j))$, bounded by a part of $\cup_{1\leq j\leq 6}(\varphi_j)$, is 3-regular, and the functional relation $w = w(u, v)$ in Ω_0 makes $w(u, v)$ a non-trivial, elliptic solution of (4.8), with parabolic degeneration in the interior of each of the boundary arcs ψ_j. It is likely, but has not been proved, that the remainder of $\partial\Omega_0$, consisting of $\cup_j(\varphi_{j-1} \cap \varphi_j)$ reduces to the six points Q_j, as suggested in Fig. 1 and 1a.

The intrinsic geometry of the closed neighborhoods $\overline{U}_j$ of each P_j is locally euclidean, with a cartesian frame consisting of the functions (u_j, u_{j+1}), while u_{j-1} becomes essentially irrelevant, subject only to the conditions of symmetry and the inequality $Lip(u_{j-1}) \leq 1$. The shape of each $\overline{U}_j$ is better suggested by Figure 1a. In terms of this geometry, the boundary arc ψ_j is a concave curve ; on the other hand, ψ_j is a geodesic in the intrinsic geometry of $\overline{\Omega}_0$; therefore $\overline{\Omega}_0$ is a simply connected disk, bounded by a closed geodesic curve: denote by a the length of one twelfth of its circumference. The intrinsic geometry of the whole hexagon Ω then looks like the visible surface of a fried egg, with $\overline{\Omega}_0$ representing the yolk, and the albumen represented by six isosceles, right triangles with right angle sides of length $1-a$, with a neighborhood of the hypotenuse hallowed out, leaving in its place a concave curve of length $2a$.

The length of each side of the hexagon Ω is $2-2a$. The geometry of the flat domains U_j implies that the positive constant a satisfies the inequalities

$$\sqrt{2} - 1 < a < \frac{1}{2} \ . \tag{6.1}$$

A preliminary calculation leads to a fairly accurate estimate of the area A of (Ω, g) of about 2.48, slightly less than the area $1.5 \cdot \sqrt{3}$ of the flat hexagon with apothem of unit length. These estimates will be applied in the next example.

7. OTHER EXAMPLES

We shall apply the example of the hexagon of the last section as a "building block" to estimate the case of an extremal isosystolic metric in a once punctured torus M. Even though this is somewhat outside of the announced scope of this paper, the complete solution of this case is important in comparing the extremal isosystolic ratios for closed surfaces of genera g and $g+1$. At this point we shall outline the example only in the case of a torus with a "sufficiently large" hole. If (M, g) is a once punctured torus, M, including its boundary circle ∂M, with a Riemannian metric g, let $\beta(M, g)$ denote the least length of any closed path in M freely homotopic to ∂M, and introduce the parameter $c = \beta(M, g)/Sys(M, g)$ to indicate the "relative size" of the puncture. Using isoperimetric inequalities of Minkowski and Mahler, one sees that the area $(Vol(M, g)$ has a positive lower bound in the space of allowable metrics, if and only if $c < 3$: the limit case, $c = 3$, yields a degenerate metric g on M, attained as a limit of metrics on the flat torus with Loewner's metric (the quotient of $\mathbb{R}^2$ by the translation lattice of a regular hexagonal tiling) by deleting increasingly large, convex, open sets from the interior of each hexagonal tile. If c is in some open interval $]c_0, 3[$, any metric g in M with relative size of the puncture equal to c and sufficiently small isosystolic ratio is an approximation of the collapsed metric of the 1-skeleton of the regular hexagonal tiling of $\mathbb{R}^2$. One can identify therefore (M, g) as the union of six narrow corridors of comparable length, each corridor meeting both others at each end, so that all three corridors intersect in two "nodes" , whose closures are disjunct ; each node meets ∂M along three disjunct arcs. Without loss of generality, we may assume that each node minimizes the total length of its boundary, subject to the constraint of meeting ∂M in three components. Then each systole-long, closed path in M runs through two corridors and both nodes. If the metric g is an extremal isosystolic one, there must be exactly three pairs of mutually opposite homotopy classes of closed paths, within which systole-long paths may occur, so that each corridor is traversed by systolic bands belonging to pairs of opposite homotopy classes. From the results of Section 2, it follows that the interior of each corridor (unless it is collapsed to a segment) is locally flat. Therefore the gradient flows of the respective potential functions must impact the boundary of the corridor at angles 45°. In fact, the two gradients are mutually perpendicular at each interior point so that at least one is transversal to the boundary of the corridor, making an angle $\alpha \neq 0$ with the generalized normal (wherever such a normal exists). In order for a systole-long path to hit the boundary of M and continue beyond as a shortest length path, it follows from Snell's laws of geometrical optics that the metric g must be discontinuous on ∂M, so that the length of any small segment σ of ∂M, each of whose points is a point of impact by some systole-long path from the interior, with a mean incidence angle α, compared to the lower limit of arc length of smooth, interior paths approaching σ from the interior of M, has ratio approaching sin α. Therefore the two systolic bands in each corridor, being mutually orthogonal in the interior of M, hit ∂M at any point where there is a normal, making equal opposite angles of $45°$, as claimed.

It follows from the above description that the nodes, when the metric is locally an extremal isosystolic one, may be obtained from the metric of the hexagon Ω described in the last section, with the following modifications. First one deletes an open neighborhood of each of the vertices P_1, P_3, P_5, in the shape of an isosceles, right triangle, with short sides of length $2-4a$; next, one rescales the arc length of the hypotenuse (where the triangle has been cut away from the hexagon Ω) by dividing it uniformly by $\sqrt{2}$, so that it too then has length $2-4a$ and becomes part of the boundary of an enneagon (polygon with nine sides) ; the intrinsic, extremal area of the resulting "fried egg" is now $A-6(1-2a)^2$, or about 2.31 ; three of its sides have, as arranged, length $2-4a$, while the other six are segments of length $2a$ in the original sides of Ω. Then one attaches to each of these last six sides of the enneagon an isosceles, right triangle with hypotenuse of the same length $2a$, so that one of its shorter sides becomes a straight continuation of the side of the enneagon, previously shortened to length $2-4a$: this continuation too becomes shortened by the factor $\sqrt{2}$, so that its new length is a. The new perimeter now consists of six straight segments, alternating each one of three sides of length $2a\sqrt{2}$ (consistent with length of nearby paths in the interior) with a side of length $2-2a$, representing a "fast track", relative to the interior, by the uniform factor $1/\sqrt{2}$. The resulting configuration of the albumen surrounding the yolk of the fried egg is now connected in the interior. This describes each of the two nodes of the surface (M,g) with the extremal isosystolic metric. The total area of each node is now $A+6a(1-a)(3a-1)$.

Finally one builds the punctured torus (M,g) by connecting the two nodes with three equal corridors, each with the following description. The interior of each corridor is isometric to a euclidean rectangle of width $2a\sqrt{2}$ and length $b\sqrt{2}$ for some arbitrary constant b. One shortens first the two sides of length $b\sqrt{2}$ of each of the three rectangles, again uniformly by the factor $\sqrt{2}$, reducing their length to b ; then one attaches one of the remaining sides of each rectangle, of length $2a\sqrt{2}$, to each of the three sides of the same length, on one of the two nodes, and the opposite, free sides similarly to the other node, matching the sides attached in such a way that the resulting surface M is orientable, and has a connected, "fast track" boundary of total length $B(M,g) = 6b + 12(1-a)$, representing its boundary systole. As a result, the manifold (M,g) has total area $Vol(M,g)$, and systole $Sys(M,g)$ expressed by the following formulas:

$$\begin{aligned} Vol(M,g) &= 2A + 12(1-a)(3a-1), \\ Sys(M,g) &= 2b + 4\ . \end{aligned} \tag{7.1}$$

If the boundary systole $\beta(M,g)/Sys(M,g) = 3 - 6a/(b+2)$ is prescribed and equal to a constant $c \in]0,3[$, this determines b uniquely in terms of the numerical constant a estimated in (6.1), at the end of the last section. This leads to a fairly accurate estimate for the extremal isosystolic ratio for M in terms of the relative length c of the boundary, provided that c is sufficently close to 3 (roughly, $c > 2.5$). While the formulas for the estimate, eventually sharpened by numerical analysis, describe what is believed to be the extremal isosystolic ratio for the punctured torus, possibly as long as $b \geq 0$ the formula cannot remain valid as c decreases indefinitely ; there must indeed be a lower limit for c for the present model of the metric to remain extremal isosystolic ; to prescribe the relative boundary systole at smaller levels would bring the nodes relatively closer to one another, until a catastrophe occurs, and the two yolks get scrambled.

The next two examples, with which we will conclude this paper, are easier: they consist of two different metrics in a closed surface M of genus 3, that are locally flat, generated by linear solutions of the Euler-Lagrange equations (5.3) for extremal isosystolic metrics ; it seems likely that they attain, locally in the function space of metrics, relative minimum values for the isosystolic ratio. Despite a careful search, no other examples of piecewise flat, extremal isosystolic metrics have been found, for surfaces of any genus ≥ 2 ; indeed there may not exist any. In both examples, the metrics admit a large, finite group of isometries ; this makes them much easier to describe.

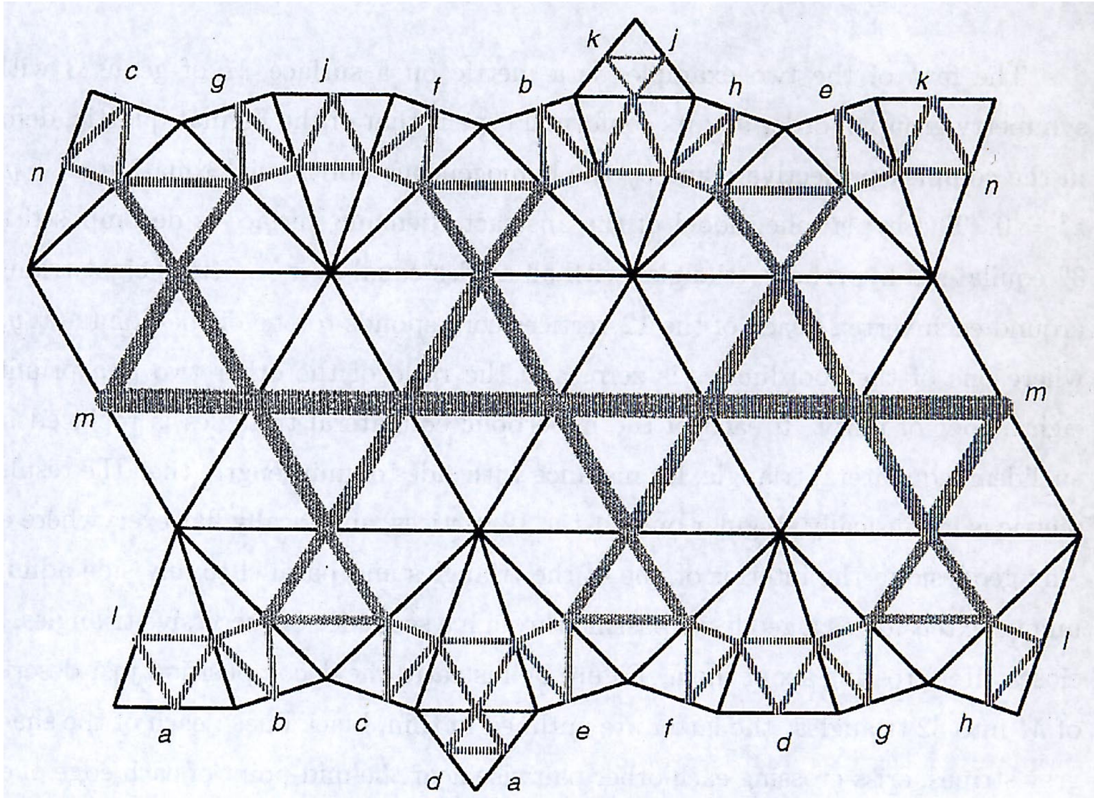


Figure 2.

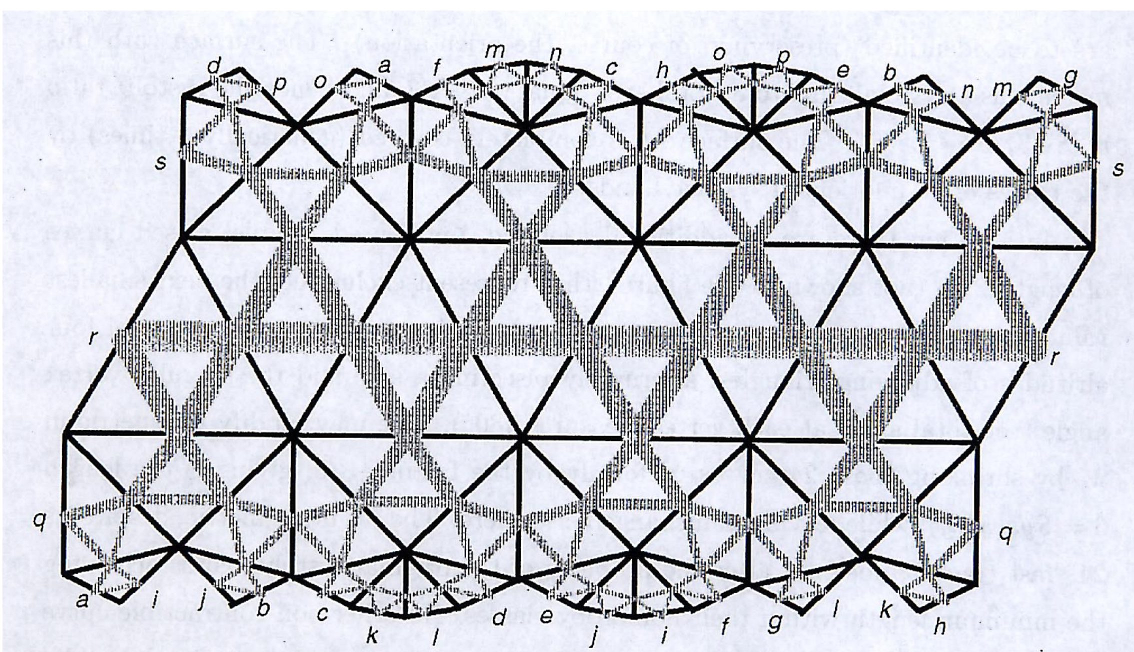


Figure 3.

The first of the two examples is a metric on a surface M of genus 3 with a symmetry group of order 96 ; its conformal type is that of the Fermat quartic, defined in the complex projective plane by the homogeneous polynomial equation $x^4 + y^4 + z^4 = 0$. The hyperbolic model of the abstract Riemann surface is decomposed into 32 equilateral hyperbolic triangles with all angles equal to $\pi/4$, fitting eight triangles around each vertex ; each of the 12 vertices corresponds to one of the points (x, y, z), where one of the coordinates is zero, and the ratio of the other two is a primitive eighth root of unity. If each of the hyperbolic equilateral triangles is replaced by a euclidean equilateral triangle, for instance with sides of unit length, then the resulting metric is intrinsically singular only at the 12 vertices, and locally flat everywhere else. Any geodesic in the interior of one of the triangles and parallel to any side admits a unique extension as a geodesic, passing through a sequence of successive triangles, and closes after crossing six of them. Figure 2 illustrates the decomposition just described of M into 32 triangles: the latter are outlined in thin, black lines ; each of the shaded, grey stripes, criss-crossing each other pairwise near the mid-point of each edge of each triangle, represents the central portion of each systolic band, and the matching pairs of letters labelling the outer edges of the diagram indicate which pairs of those edges are to be identified (preserving, of course, the orientation). The surface with this metric has a systole equal to 3, total area equal to $8 \cdot \sqrt{3}$, and hence the systolic ratio is $(8\sqrt{3})/9 \sim 1.5396$. The surface M is completely covered (generically 3 times) by the traces of 16 unoriented systolic bands.

In addition there are 24 additional, isolated, unoriented, simple, closed curves of length $2\sqrt{3}$ (not shown in the figure), that represent exclusively the next smallest value attained by local systoles. Each of them consists of a cyclic sequence of four altitudes of adjoining triangles, alternately bisecting a side and the singular vertex angle (the total angle at each vertes measures 480°). One may modify the metric in M by shrinking these 24 curves uniformly by the factor $\frac{1}{2}\sqrt{3}$, giving it the length $3 = Sys(M, g)$, while leaving g unchanged elsewhere. The resulting metric, despite its 24 "fast tracks", does not alter the property of the original systolic bands achieving the minimum length within their homotopy classes. No other non contractible curve in M may be shrunk to one of length 3 without causing some of the original systole-long curves to cease being length-minimizing in their class. Therefore M , with the metric described above, and altered by shrinking the 24 curves down to length 3 has a maximal set of critical isosystolic classes consisting of 80 free homotopy classes, out of which 32 make up the essential critical classes, while the other 48 classes systolic bands consisting of just one curve each.

The second example is conformally equivalent to Felix Klein's sextic curve and has a symmetry group isomorphic to $PSL\,(2, \mathbb{Z}/7\mathbb{Z})$, of order 168 (it is the well known example achieving A. Hurwitz's upper bound for the order of the group of symmetries of such a surface). It corresponds, as a Riemann surface, to the compactification of the quotient of Poincaré's upper half-plane by the free action of the congruence subgroup

$$\Gamma_y - \mathrm{Ker}\{(PSL(2, \mathbb{Z}) \to PSLL(2, \mathbb{Z}/7\mathbb{Z})\} \subset PSL(2, \mathbb{Z})\ ;$$

the compactification involves adjoining 24 points, to form a compact Riemann surface of genus 3. The hyperbolic metric in the resulting, compact surface consists of 56 hyperbolic, equilateral triangles with angles measuring $2\pi/7$, whose vertices are the 24 points adjoined in the compactification. As before, each of these 56 triangles is replaced by a euclidean equilateral triangle of unit side length. The resulting surface, illustrated diagrammatically in Figure 3, if locally flat, like the one in the previous example, except for the 24 singular points. The systole is achieved by any geodesic path parallel to any side of any triangle ; its completion as a geodesic always closes after crossing 8 triangles (see Fig. 3). Thus the area of M with this metric is $14 \cdot \sqrt{3}$, the systole has length 4, and therefore the systolic ratio is $(7\sqrt{3})/8 \sim 1.5155$, or about 1.5% lower than the one for the previous metric. It seems very likely that the systolic ratio $(7\sqrt{3})/8$ achieved by this metric is the absolute minimum value for a surface of genus 3, or else is probably very close to the minimum. With this metric, the surface is covered almost everywhere three times by 21 bands of unoriented systole-long paths. The next smallest value of a local systole is $3\sqrt{3}$, achieved by 28 bands of unoriented closed curves (not shown in the figure), also covering M, generically 3 times: The excess of this "second" systole over the principal value 4 is too great, to permit any "fast track" shrinking of curves in any other non-trivial, free homotopy class down to length 4, without destroying the minimal length property of the original systole-long curves.

It may be interesting to note that the conformal equivalence classes of the metrics described in the two above examples are quite far apart in the space of moduli of surfaces of genus 3. Assuming that both metrics represent relative minimizing points for the isosystolic ratio, this indicates that the isosystolic ratio, as a functional on the space of all metrics in a surface M, is not a convex variational functional.

It was mentioned in the introduction that it seems unlikely that, for surfaces of genus $g \geq 2$ and $\neq 3$, the extremal isosystolic metrics may be piecewise flat, as it is in the last two examples ; in the case $g = 2$ the non-existence of a piecewise flat, extremal isosystolic metric may be shown by a reduction to a symmetry argument. Some attempts, so far without success, have been made in the case $g = 4$, using metrics with a symmetry group of order 120.

In conclusion, I wish to recall that a great deal of what I originally learned about the isosystolic problem has been from lectures by Marcel Berger. For this and for many other fruitful ideas that conversations with him have suggested, I wish to express my appreciation, and thereby end this exposé.

BIBLIOGRAPHY

[1] C. Bavard, *Inégalité isosystolique pour la bouteille de Klein,* Math. Ann. **274** (1986), 439–441.

[2] C. Bavard, *Inégalités isosystoliques conformes pour la bouteille de Klein,* Geometriae Dedicata **27** (1988), 349–355.

[3] M. Berger, *Du côté de chez Pu,* Ann. Sci. Ecole Norm. Sup. Paris **5** (1972), 1–44.

[4] M. Berger, *A l'ombre de Loewner,* Ann. Sci. Ecole Norm. Sup. Paris **5** (1972), 241–260.

[5] E. Cartan, *Les espaces métriques fondés sur la notion d'aire,* Exposés de Géométrie, I, Paris, Hermann (1933).

[6] M. Gromov, *Filling Riemannian Manifolds,* J. Differential Geom. **18** (1983), 1–147.

[7] P.M. Pu, *Some inequalities in certain nonorientable Riemannian manifolds,* Pac. Jour. Math. **2** (1952), 55–71.