A topological proof of the Arnold four cusps theorem

V.A. VASSILIEV

To the memory of V.I. Arnold

Abstract. The Arnold's theorem (generalizing a consideration by Jacobi) states that on a generic Riemannian surface, which is sufficiently close to a sphere, the k-th caustic of a generic point has at least four semi-cubical vertices. We prove this fact by the methods of the Morse theory; in particular we replace the previous analytical condition of the "sufficient closeness to the sphere" by a geometric one, which probably is considerably less restrictive.

Let M^2 be a compact smooth Riemannian 2-dimensional manifold, $p \in M^2$, and $\Phi : (T_pM^2, 0) \to (M^2, p)$ the geodesic map, sending any central ray of T_pM^2 into the geodesic line, passing from p to the corresponding direction so that in the restriction to this ray the map Φ is isometric. The *k-th caustic* $C_k(p)$ of the point p is the union of k-th conjugate points on all these geodesic lines, see. [6], [3].

Let the Riemannian metric on M^2 be elliptic and generic, then for any k $C_k(p)$ is a compact curve in M^2, all whose singularities are transversal self-intersections and semicubical cusps only. Denote by $\tilde{C}(p) \subset T_pM^2$ the union of critical points of the geodesic map Φ. This set splits into the union of curves $\tilde{C}_k(p)$, homeomorphic to circles and consisting of the k-th intersection points of our rays with $\tilde{C}(p)$. Then $C_k(p) \equiv \Phi(\tilde{C}_k(p))$; this map has the *fold singularity* over the non-singular points of the caustic, and the *Whitney cusp* singularities over the cusps, see [4], §3.1. In addition, in T_pM^2 the norm function is defined by our metric, and the cusps of $C_k(p)$ are exactly the images of extrema of the restriction of this function to $\tilde{C}_k(p)$. Any locally non-singular branch of the caustic has the standard *co-orientation* (i.e. orientation of normal directions): to the side, on which the geodesic map locally has more pre-images.

The standard unit sphere S^2 is an example of a non-generic surface: any its set $C_k(p)$ is not a curve but, depending on k, either p itself or its opposite point. Nevertheless all the critical sets $\tilde{C}_k(p)$ are still non-singular: they are concentric circles of radii πk. C.-G. Jacobi has noticed that on a generic ellipsoid the first caustic of a generic point always has at least four cusps. V.I. Arnold [1], [2], using some ideas of S.L. Tabachnikov [7], has proved the following theorem.

Theorem 1. *For any k and any generic surface, sufficiently C^∞-close to the standard sphere, the k-th caustic has at least 4 cusps (where the condition of closeness to the sphere strengthens when k grows).*

The proof in [1] is analytic. Below we give a topological proof. The conditions of "closeness to the sphere" used in it also are topological and probably much less

Date: 8th June 2011.

Supported in part by the grant NSh-8462.2010.1 of the President of Russia for the support of leading scientific schools.

restrictive than the ones in [1]. Let us formulate them. A weakly deformed sphere S^2 can be identified with the initial one via the orthogonal projection, therefore instead of the deformation of the sphere we will consider the corresponding deformation of the Riemannian metric on the fixed sphere. Let us denote by S^1_T the space of rays in T_pS^2 with the origin at 0.

Definition 1. A Riemannian metric in S^2 (and the corresponding system of geodesics) is *k-close* to the standard metric of the sphere, if it is elliptic, and there is a closed embedded disc $D \subset S^2$, containing the k-th caustic $C_k(p)$ and such that

(1) any two points of D can be connected in D by a unique geodesic segment;
(2) for any ray $c \subset T_pS^2$ with the origin at 0, the connected component of the set $\Phi^{-1}(D) \cap c$, intersecting the curve $\tilde{C}_k(p)$, does not meet other curves $\tilde{C}_m(p)$, $m \neq k$;
(3) two maps $S^1_T \to \partial D$, associating any such ray c with the images of endpoints of this connected components under the map Φ, are diffeomorphisms;
(4) let $\bar{\pi}$ be the maximal length (in our Riemannian metric) of meridians connecting the point p with its antipodal point $-p$; let r be the maximal geodesic distance from the points of the disc D to p if k is even or to $-p$ if k is odd; let Δ be the minimal norm in T_pS^2 of points of the connected components of the set $\Phi^{-1}(D)$, which are exterior to the component connecting $\tilde{C}_k(p)$; then the inequality $k\bar{\pi} + r < \Delta$ should hold.

Example 1. If the metric is standard, and D is the disc of radius r centered at the point $(-1)^k p$, then $\bar{\pi} = \pi$, $\Delta = (k+2)\pi - r$. In this case all conditions of Definition 1 are satisfied for any $r \in (0, \pi/2)$.

Below, we prove Theorem 1 in exactly this understanding of the words "sufficiently close to the standard sphere". Let us start with an obvious property of caustics. Introduce a cyclic coordinate $\alpha \in \mathbb{R}^1/2\pi\mathbb{Z}$ on ∂D. Given a tangent element in D (i.e. a couple consisting of a point and a tangent direction at it), its *bending angle* is half the sum of coordinates of endpoints of the geodesic segment in D, tangent to this tangent element. This angle is a point of the circle $\mathbb{R}^1/\pi\mathbb{Z}$; by condition (1) it is always well-defined. For a generic caustic $C_k(p)$, its tangent elements are defined not only at its non-singular points, but also at the cusps.

Proposition 1. *If the conditions of Definition 1 are satisfied, then the k-th caustic $C_k(p)$ has a surjective parameterization $S^1 \to C_k(p)$ such that the bending angle of the tangent element of the caustic grows monotonically over the parameter, and when this parameter passes once the entire circle S^1, this angle turns exactly two times along the circle $\mathbb{R}^1/\pi\mathbb{Z}$.*

Proof. This parameterization sends any point of the circle S^1_T to the k-th conjugate to p point of the corresponding geodesic. By condition (3) of Definition 1, when c moves along S^1_T, both endpoints of corresponding geodesic segments move monotonically and accomplish one rotation along the circle $\mathbb{R}^1/2\pi\mathbb{Z}$. □

If a generic caustic has less than 4 cusps, then there are 2 of them: the images of absolute extremal points of the norm function on the curve $\tilde{C}_k(p) \subset T_pS^2$.

Proposition 2. *In the conditions of Definition 1, there is only one* (*up to isotopy*) *curve with 2 cusps in D, having no singularities except for transversal self-intersections and these cusps, and admitting a parameterization satisfying the assertions*

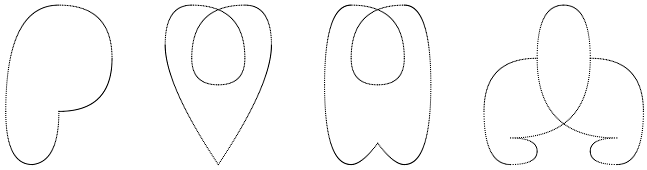
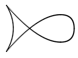


Figure 1. Forbidden fragments of the caustic

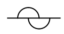
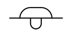
of Proposition 1: namely, it is the curve  *(well-known in this theory, see e.g.* [2], [8]*).*

Proof. Consider some curve with exactly two cusps, satisfying these conditions; choose one of two branches of this curve, connecting these two cusps. Its self-intersection points define a *chord diagram* on the segment parameterizing this branch: it is obtained from this segment by adding the arcs, connecting the pre-images of any such point, cf. [9].

Lemma 1. *The rotation of the bending angle on the segment between the endpoints of any chord is greater than the complete twist along* $\mathbb{R}^1/\pi\mathbb{Z}$.

Proof. Let K be the image of this segment: it is a closed immersed curve in D with only one breakpoint. For any point $a \in \partial D$ denote by $t(a, K)$ the number of geodesic segments in D with an endpoint at a, that are tangent to K at some its non-singular point. Then the rotation of the bending angle along K equals half the integral of the form $t(a(\alpha), K)d\alpha$ along ∂D. The value of the function $t(a, K)$ is nowhere less than 1. Indeed, by condition (1) of Definition 1 the set of geodesics, going from $a \in \partial D$ inside D and intersecting K, has at least two boundary points; at most one of them can not to be a geodesic tangent to K. Moreover, there obviously exist points a, at which $t(a, K) \geq 2$. So, our integral is greater than π. $\square$

Now let us go from some cusp point along the chosen branch of our curve until the instant when this branch intersects for the second time its already passed part (if such an instant exists). If the corresponding two arcs are not linked in this branch (i.e. both endpoints of one arc are placed before both endpoints of the other), then by Lemma 1 the bending angle makes more than two complete rotations already on the two (separated) segments between these endpoints, in contradiction to the condition.

Lemma 2. *If these two arcs are linked by the type*  *or*  *, then the bending angle of our branch of the curve* $C_k(p)$ *makes more than two complete rotations already on the segment between endpoints of these arcs.*

Proof. In the first case our curve has a piece K, isotopic to one shown in the left-hand part of Fig. 1. The function $t(a, K)$, defined as in the proof of Lemma 1, in this case is nowhere less than 2: both boundary points of the set of geodesics defined there should be tangent to K. In the second case we have a fragment isotopic to one of two curves shown in the central part of Fig. 1. Then both the interior and the exterior parts of this curve provide the rotations of the bending angle, greater than the complete one: for the interior part of the curve this follows immediately from Lemma 1, and for the exterior one from a very similar argument. $\square$

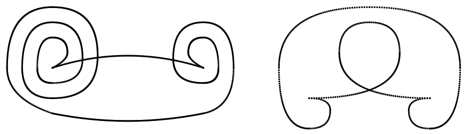
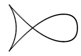


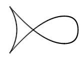
FIGURE 2. Curves with monotone bending angle

So, none of two branches of $C_k(p)$, connecting two its cusp points, can have two self-intersections, and at least one of them has no one. There are only two possibilities remained: either we have no self-intersections of these branches, and only intersections of one branch with the other, or additionally there is a single self-intersection of one of branches.

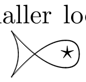
In the first case the topological types of the mutual intersections of these branches, not contradicting the monotonicity of the bending angle, are depicted in the left-hand part of Fig. 2 with arbitrary (maybe equal to zero) numbers of intersections on the left and on the right. In any of these cases already the exterior contour of our curve provides more than two complete rotations of the bending angle.

In the second case suppose that there is at least one intersection point of these branches. If at least one of corresponding points of the self-intersecting branch is placed in it outside of the segment between the pre-images of the self-intersection points of this branch, then the curve $C_k(p)$ contains a fragment isotopic to the figure  or  (in which the angle of the upper wedge is equal to zero, while that of the lower one is not). Exactly as in Lemma 1 we obtain that the rotation of the bending angle along this fragment in the first case is greater than π, and in the second one greater than 2π; in addition by Lemma 1 the segment between the self-intersection points contributes more than π. If the self-intersecting branch is intersected by the other one only between its self-intersection points, then $C_k(p)$ contains a fragment isotopic to the one shown in Fig. 1 on the right or on the left. Already the rotation of the bending angle provided by any of these fragments is greater than 2π.

Finally, only one possibility has remained: our branches have no mutual intersection points, and exactly one of them has a single self-intersection. There are only two isotopy types of such curves:  and one shown in the right-hand part of Fig. 2. The same arguments as before show that the latter picture provides the rotation number greater than 2, and Proposition 2 is proved. □

Proof of Theorem 1. Suppose that the caustic $C_k(p) \subset D$ is isotopic to  . Let $\tilde{D} \subset T_pS^2$ be the connected component of the set $\Phi^{-1}(D)$, containing $\tilde{C}_k(p)$. By condition (2) of Definition 1, any point $x \in \partial D$ has exactly two pre-images under the map

$$\Phi|_{\tilde{D}} : \tilde{D} \to D. \qquad (1)$$

When x crosses $C_k(p)$ in the direction of the standard co-orientation, the number of its pre-images increases by 2. Close to cusp points this co-orientation is always directed inside the smaller local area of the complement of the caustic, hence in the right-hand part of  it is directed outside. Therefore the points from the right-hand (containing the point $\star$) component of the figure bounded by this curve have no pre-images under the map (1); this means that no geodesics from p enter this component close to their k-th conjugate points. After that any geodesic leaves the disc D and comes back to it at some point, whose distance from p along this geodesic is not less than $\Delta > k\bar{\pi} + r$, see condition (4) of Definition 1. So, any geodesic, connecting the points $\star$ and p and containing at least k points conjugate to p, is no shorter than Δ. For $k = 1$ this contradicts the Hopf-Rinow theorem [5], [6]: the distance from any point of the disc D to p is no more than $\bar{\pi} + r$, hence there is a geodesic connecting them, whose length does not exceed this number.

Now let be $k > 1$. Choose a point $\star$ in our component, which is not conjugate to p along any geodesic of length $\leq 2k\bar{\pi} + r$. Consider the space $\Omega(p, \star; k\bar{\pi} + r)$ of all piecewise smooth paths from p to $\star$ of length $\leq k\bar{\pi} + r$. This space has non-trivial homology groups in all dimensions $0, 1, \dots, k$; moreover, the identical embedding of this space into the entire space $\Omega(p, \star)$ of all continuous paths from p to $\star$ induces non-trivial homology maps in all these dimensions. Namely, for any $i = 1, \dots, k$ a non-trivial i-dimensional cycle in $\Omega(p, \star)$ is realized by the fundamental class of the imbedded i-dimensional torus, whose points are piecewise smooth curves, consisting of $(i+1)$ smooth segments, the first of which is a fixed path of length $\leq r$ (if $i = k$) or $\leq \bar{\pi} + r$ (if $i < k$) from $\star$ to $(-1)^i p$, and remaining i segments are arbitrary meridians connecting p and $-p$.

In particular, the energy function (i.e. simply the length of paths) on the space $\Omega(p, \star; k\bar{\pi} + r)$ has critical points of any integer index $i \in [0, k]$. By [6], Theorem 16.2, such critical points are nothing else than the geodesics from p to $\star$ with exactly i conjugate points. A contradiction. $\square$

References

[1] V.I. Arnold, Sur les propriétés topologiques des projections Lagrangiennes en geométrie symplectique des caustiques. Preprint no. 9320 CEREMADE, Cahiers Mathematiques de la Decision, Université Paris-Dauphine. 14/6/93, 1-9. Russian transl. in: V.I. Arnold, Selecta – 60, Phasis, Moscow, 1997; 525–532.

[2] V.I. Arnold, Topological Invariants of Plane Curves and Caustics. AMS, University Lecture series, vol. 5, 60 pp.

[3] V.I. Arnold. Singularities of Caustics and Wave Fronts. Kluwer, Dorderecht-Boston-London, 1990.

[4] V.I. Arnold, V.V. Goryunov, O.V. Lyashko, V.A. Vasil'ev. Singularity Theory I. 2nd ed. Springer, Berlin a.o., 1998, 245 pp.

[5] H. Hopf, W. Rinow. Über den Begriff der vollständigen differentialgeometrischen Fläche. Comm. Math. Helv. 3 (1931), 209–225.

[6] J. Milnor, Morse Theory. Princeton Univ. Press, Princeton NJ, 1963.

[7] S.L. Tabachnikov, Around four vertices. Russian Math. Surveys, 45:1 (1990), 229-230.

[8] S.L. Tabachnikov, The four-vertex theorem revisited – two variations on the old theme. American Mathematical Monthly, 102:10 (1995), 912–916.

[9] V.A. Vassiliev, Complements of discriminants of smooth maps: topology and applications. Revised ed. Thanslations of Math. Monographs, 98. AMS, Providence RI., 1994, 266+ix pp.

Steklov Mathematical Institute and Higher School of Economics, Moscow, Russia
E-mail address: vva@mi.ras.ru