DIFFERENTIABLE ISOTOPIES ON THE 2-SPHERE

James Munkres

Let Diff S^n denote the group of diffeomorphisms of degree +1 and of class C^r $(1 \leq r \leq \infty)$ carrying the unit n-sphere onto itself, topologized by requiring closeness of the maps and their partial derivatives through order r. A path in this space is called a *regular isotopy;* it is a map of the reals R into the space which is constant on the set $t \leq 0$ and on the set $t \geq 1$. We say a path is *differentiable* if the induced map of $S^n \times R$ onto S^n is of class C^1; a differentiable path is also called a *differentiable isotopy.* If two maps are regularly isotopic, they are differentiably isotopic as well [2, Lemma 1.6].

The group Γ^{n+1} of Milnor and Thom is defined as a quotient group of the group $\pi_0(\text{Diff } S^n)$ of path components of this space. The object of the present paper is to give an elementary proof that $\pi_0(\text{Diff } S^2)$, and hence Γ^3, vanishes. This result has since been generalized by Smale [3]. Interest in the group Γ^{n+1} stems from its close connection with the existence of distinct differentiable structures on manifolds; the fact that $\Gamma^3 = 0$, for example, implies the uniqueness theorem for differentiable structures on 3-manifolds [2]. The group $\pi_0(\text{Diff } S^n)$ does not in fact depend on the choice of r; for simplicity we shall prove our result only in the case $r = 1$.

1. Let $\mathcal{E}^r(R^m, R^n)$ $(m \leq n)$ denote the space of those embeddings of class C^r of R^m in R^n which equal the inclusion map i outside some compact subset of R^m; it is topologized as was Diff S^n. (The map i is defined by the equation

$$i(x_1, x_2, \cdots, x_m) = (x_1, \cdots, x_m, 0, \cdots, 0).)$$

A *loop* in this space will be assumed to be based at i and to be constant for t near 0 and for t near 1. Every element ϕ of Diff S^n is regularly isotopic to an element ϕ_1 which equals the identity in a neighborhood of the north pole [2, Lemma 8.1]; stereographic projection carries ϕ_1 into an element f of $\mathcal{E}^r(R^n, R^n)$. A path in $\mathcal{E}^r(R^n, R^n)$ is carried by the inverse of this projection into a path in Diff S^n; hence our problem reduces to showing that $\pi_0(\mathcal{E}^1(R^2, R^2)) = 0$.

1.1. LEMMA. *Let* f_t *be a differentiable loop in* $\mathcal{E}^2(R^1, R^2)$ *which is homotopic to a loop in the subspace* $\mathcal{E}^2(R^1, R^1)$, *the homotopy passing through differentiable loops. Then there exists a differentiable loop* G_t *in* $\mathcal{E}^1(R^2, R^2)$ *such that for each* t, $G_t f_t$ *maps* R^1 *into itself.*

Proof. Given $g \in \mathcal{E}^2(R^1, R^2)$, there exists a neighborhood of g such that if h lies in this neighborhood, there exists a C^1 diffeomorphism J of R^2 which carries the set $g(R^1)$ onto $h(R^1)$. The diffeomorphism is obtained by sliding $g(R^1)$ along its normal lines onto $h(R^1)$ and leaving everything fixed outside a neighborhood of these sets. Because the maps are of class C^2, there is no difficulty in carrying out this construction; details are left to the reader. Similarly, if h_t is a differentiable loop approximating closely enough the differentiable loop g_t, then the diffeomorphism J_t may be chosen to be a differentiable loop in $\mathcal{E}^1(R^2, R^2)$.

Received July 31, 1959; in revised form, June 23, 1960.

Presented to the American Mathematical Society August 26, 1958. This work was supported by contract AF 18(600)-1494.

Since f_t is homotopic to a loop in $\mathcal{E}^2(R^1, R^1)$, there exist a finite number of such diffeomorphisms J_t such that their composition throws $f_t(R^1)$ onto R^1 for each t. This composition is the required loop G_t.

1.2. LEMMA. *Let* G_t *be a differentiable path in* $\mathcal{E}^1(R^2, R^2)$ *such that the map* G_t *carries* R^1 *into itself for each* t, *and equals the identity on the intersection of the sets* $|y| < \varepsilon$ *and* $|t - 1/2| > 1/2 - \varepsilon$. *Then there exists a differentiable loop* H_t *in* $\mathcal{E}^1(R^2, R^2)$ *such that, for each* t, $H_t G_t$ *equals the identity in a neighborhood of* R^1.

Proof. A differentiable path G_t in $\mathcal{E}^1(R^2, R^2)$ induces a C^1 diffeomorphism $g(x, y, t) = (G_t(x, y), t)$ of R^3 onto itself which carries $R^2 \times t$ onto itself for each t, and conversely. Let b be so chosen that $G_t(x, y) = (x, y)$ except for $|x| < b/2$. Let M be the subset of R^3 on which $|x| \leq b$, $y = 0$, and $0 \leq t \leq 1$. Consider the restriction of g first to the half-space $y \geq 0$ and then to the half-space $y \leq 0$. Apply 8.3 and 8.4 of [2] with M as just defined and with A a neighborhood of Bd M. We obtain a diffeomorphism $g_1(x, y, t)$, defined whenever $(x, 0, t)$ is in M, such that

$$g_1(x, y, t) = g(x, 0, t) + (0, y, 0)$$

for $|y| < \delta$ (for some δ). We require g_1 to equal g except for $|x| < 2b/3$, $|y| < \varepsilon/2$, and $|t - 1/2| < (1 - \varepsilon)/2$, so that g_1 may be extended to R^3 by defining it equal to g elsewhere. One must go to the proof of 8.3 of [2] to verify that g_1 maps $R^2 \times t$ into itself for each t.

Now let $\beta(y)$ be a C^∞ function which equals 1 for y near 0 and equals 0 for $|y| \geq \delta/2$. Set $g_2(x, y, t) = g_1(x, y, t)$ for $|y| > \delta/2$ and

$$g_2(x, y, t) = \beta(y)(x, 0, t) + (1 - \beta(y))\, g(x, 0, t) + (0, y, 0)$$

for $|y| < \delta$. Then g_2 is a diffeomorphism which maps $R^2 \times t$ onto itself for each t, and $g_2(x, y, t) = (x, y, t)$ for y near 0. Set $h = g_2 g^{-1}$; the loop H_t in $\mathcal{E}^1(R^2, R^2)$ induced by h satisfies the demands of the lemma.

1.3. THEOREM. *If* $f \in \mathcal{E}^1(R^2, R^2)$, *then* f *is regularly isotopic to the identity.*

Proof. We may assume that $f \in \mathcal{E}^2(R^2, R^2)$, since

$$\tilde{f}(x, y) = \frac{1}{4\delta^2} \int_{x-\delta}^{x+\delta} \int_{y-\delta}^{y+\delta} f(s, t)\, dt\, ds$$

is in $\mathcal{E}^2(R^2, R^2)$ for small δ, and is regularly isotopic to f (as $\delta \to 0$, $\tilde{f}$ converges to f in the C^1-topology). We may also assume that f equals the identity in a neighborhood of $y \leq 0$. Choose b so that f equals the identity for $y \geq b/2$. Let $\alpha(t)$ be a C^∞ function with $\alpha(t) = 0$ for $t \leq 1/3$, $\alpha(t) = 1$ for $t \geq 2/3$, and $\alpha'(t) \geq 0$.

Define a differentiable path F_t in $\mathcal{E}^2(R^2, R^2)$ by the equation

$$F_t(x, y) = f(x, y + \alpha(t)b) - (0, \alpha(t)b).$$

Now F_t equals the identity in a neighborhood of R^1 for t near 0 and for t near 1. Hence $f_t(x) = F_t(x, 0)$ is a differentiable loop in $\mathcal{E}^2(R^1, R^2)$. We shall prove (2.1) that the loop f_t is homotopic to a loop in $\mathcal{E}^2(R^1, R^1)$, the homotopy passing through differentiable loops. Let G_t be the loop constructed in 1.1; then $G_t F_t$ is a differentiable path in $\mathcal{E}^1(R^2, R^2)$ which maps R^1 into itself for each t and equals the identity in a neighborhood of R^1 for t near 0 or 1. By 1.2, there exists a differentiable loop H_t in $\mathcal{E}^1(R^2, R^2)$ such that $H_t G_t F_t$ is the identity in a neighborhood of R^1 for each t.

Define $h_t(x, y)$ equal to $H_t G_t F_t$ for $y \geq 0$ and equal to the identity for $y \leq 0$. Then h_t is a differentiable path in $\mathcal{E}^1(R^2, \overline{R^2})$. For t near 0, $F_t(x, y) = f(x, y)$, so that $h_0(x, y) = f(x, y)$. For t near 1 and $y \geq 0$,

$$F_t(x, y) = f(x, y + b) - (0, b) = (x, y + b) - (0, b) = (x, y).$$

Hence $h_1(x, y) = (x, y)$ for all (x, y).

2. One may ask where these arguments would break down if we attempted to prove that $\pi_0(\mathcal{E}^1(R^n, R^n)) = 0$ (which would imply that $\Gamma^{n+1} = 0$). One such point appears in the second paragraph of 1.2, where one would need to have

$$\pi_1(\mathcal{E}^1(R^{n-1}, R^{n-1})) = 0.$$

(This is trivial in the case $n = 2$.) The other difficulty appears if we attempt to generalize to higher dimensions the following lemma, whose proof occupies the remainder of the paper.

2.1. LEMMA. *A differentiable loop* f_t *in* $\mathcal{E}^2(R^1, R^2)$ *is homotopic to a loop in* $\mathcal{E}^2(R^1, R^1)$, *the homotopy passing through differentiable loops.*

2.2. *Definition.* Let G be a polygonal curve in R^2. A vertex q of G is said to be *admissible* if it is not an end point of G and if the triangle based at q (that is, the closed convex hull of p, q, r, where p and r are the vertices adjacent to q) intersects G only in the edges pq and qr.

2.3. LEMMA. *Let* G *be a simple closed polygonal curve in* R^2 *with more than three vertices; let* σ *be an edge of* G. *There exists an admissible vertex* r *such that the triangle based at* r *is contained in* Cl(Int G), *and* r *is not a vertex of* σ.

Proof (compare [1]). We proceed by induction on the number n of edges in G. Let s, t be the vertices of σ. The sum of the interior angles of G is $(n - 2)\pi$; hence there exists a vertex $p \neq s, t$ whose interior angle is less than π. If p is admissible, the lemma holds. Otherwise, there exists a vertex q of G in the triangle based at p, such that the open segment pq lies in Int G. $G \cup (pq)$ falls into two simple closed curves, with disjoint interiors, having only the edge pq in common. One of these curves contains σ; application of the induction hypothesis (with pq taking the place of σ) to the other curve locates the required admissible vertex.

2.4. LEMMA. *Let* G *be a simple polygonal curve in* R^2 *lying in the region* $a < x < b$ *except for its end points* (a, 0) *and* (b, 0). *If* p_i *is an inadmissible interior vertex of* G, *then* G *has an admissible vertex not adjacent to* p_i.

Proof. Let $(a, 0) = p_0, p_1, \cdots, p_n = (b, 0)$ be the successive vertices of G. For each inadmissible vertex $p_j \neq p_0, p_n$ of G, define $g(p_j)$ as follows: Consider the vertices of G lying in the triangle based at p_j, other than p_{j-1} and p_{j+1}. Consider lines passing through these vertices and parallel to $p_{j-1}p_{j+1}$; pick $g(p_j)$ as a vertex on the line closest to p_j. Let G' denote the polygonal curve with successive vertices $p_i, g(p_i), g^2(p_i), \cdots$. There are several possibilities: G' may end at a vertex other than p_{i-1} or p_{i+1}, in which case the lemma follows; G' may end at p_{i-1} or p_{i+1}; or we may have $g^j(p_i) = g^k(p_i)$ for some $j < k$. Consider the last case; let k be the smallest such integer; let G" be the closed polygonal curve with successive vertices $g^j(p_i), g^{j+1}(p_i), \cdots, g^{k-1}(p_i)$. One sees readily that G" is simple, and that it crosses G at each of the vertices of G" but is otherwise disjoint from G. Consider the components of $G \cap$ Int G"; among these arcs are two that join adjacent vertices of G". At least one of these two arcs contains neither p_{i-1} nor p_{i+1} as an interior point; application of Lemma 2.3 to the union of this arc with the corresponding line segment of

G'' (with σ set equal to this line segment) locates the desired admissible vertex. A similar argument applies in the other case; the fact that G' ends at a vertex adjacent to p_i on G is essential.

2.5. Let $\mathfrak{P}_n(R^2)$ denote the space of all simple polygonal curves (graphs) G in R^2 from $(a, 0)$ to $(b, 0)$, lying in $a < x < b$ except for these two end points, and having n vertices ($n \geq 3$). The space is topologized by requiring closeness of corresponding vertices; a path G_t in this space is *differentiable* if the vertices are C^1 functions of t. $\mathfrak{P}_n(R^1)$ is the subspace of graphs which lie in R^1.

2.6. LEMMA. *Let G_t be a differentiable path in $\mathfrak{P}_n(R^2)$. Corresponding to any $\delta > 0$ and $0 < t_0 < 1$, there is a homotopy of G_t to a path H_t such that H_t lies in $\mathfrak{P}_n(R^1)$ for t near t_0. The homotopy passes through differentiable paths, and it is the identity for $|t - t_0| > \delta$.*

Proof. We proceed by induction on n. Let p be an admissible vertex of G_{t_0}; it is then also an admissible vertex for all G_t with $|t - t_0| < \delta_1 < \delta$ (for some such δ_1). A homotopy is constructed by gradually pushing in the triangle based at p, for those G with $|t - t_0| < \delta_1/2$, so that the edges incident on p make a straight angle. For $\delta_1/2 < |t - t_0| < \delta_1$, one pushes the triangle in only part-way, and for $|t - t_0| > \delta_1$ one does nothing at all.

Now that portion of the resulting path which corresponds to the interval $|t - t_0| \leq \delta_1/3$ may be considered as an element of $\mathfrak{P}_{n-1}(R^2)$, if we delete p as a vertex, and the induction hypothesis may be applied to obtain a homotopy. We require it to be the identity for $|t - t_0| > \delta_1/4$, so that it may be extended to the entire t-interval by making it the identity outside. The composition of these two homotopies satisfies the demands of the lemma.

2.7. LEMMA. *Let G_t be (1) a differentiable path in $\mathfrak{P}_n(R^2)$ which (2) is constant and lies in $\mathfrak{P}_n(R^1)$ for t near 0 and for t near 1. There is a homotopy of G_t to a path in $\mathfrak{P}_n(R^1)$, the homotopy passing through paths satisfying (1) and (2).*

Proof. Again we proceed by induction. Using 2.4, we may construct a subdivision $t_0, \cdots, t_m$ of the t-interval and a correspondence between the intervals $I_i = [t_{i-1}, t_i]$ and the vertices of G_t, such that for each i the vertex p^i corresponding to I_i is admissible in G_t for all t in a neighborhood of I_i, and p^i and p^{i+1} are distinct and non-adjacent. As in 2.6, we push in the triangle based at p^1 so that the angle at p^1 is straight, for those G_t with t in a neighborhood of I_1. Near $t = t_0 = 0$ there is no pushing in, since the angle at p^1 is already straight; for $t > t_1$ we taper off the "pushing in" sharply, so that the homotopy thus defined is the identity except in a small neighborhood of I_1. Because p^2 is not adjacent to p^1, its admissibility for $t \in I_2$ is not affected by the alteration we just carried out in G_t. For the same reason, a similar pushing in of the triangle based at p^2, for all $t \in I_2$, will not affect the straightness of the angle at p^1 for $t \in I_1$ (as it would, of course, if p^2 were adjacent to p^1). We proceed in this way for each successive interval. Let the resulting path be denoted by H_t.

Now we consider the interval $|t - t_i| < \delta$, where δ is chosen so that the angles at p^i and p^{i+1} are straight for these values of t. We then delete p^i and p^{i+1} as vertices, consider H_t as a graph in $\mathfrak{P}_{n-2}(R^2)$, and apply 2.6: There is a homotopy of H_t to a path which lies in $\mathfrak{P}_{n-2}(R^1)$ for $|t - t_i| < \delta_i < \delta/2$; the homotopy is required to be the identity for $|t - t_i| > \delta/2$, so that it may be extended trivially to the entire t-interval. Because we considered H_t as lying in $\mathfrak{P}_{n-2}$ rather than $\mathfrak{P}_n$, the angles at p^i and p^{i+1} remain straight for $|t - t_i| < \delta$.

We proceed in the same way for each i. Let the resulting path be denoted by J_t. For $t \in I_i$, the angle at p^i is flat, so that J_t may be considered as a path in $\mathfrak{P}_{n-1}(R^2)$, on deletion of p^i as a vertex. Since J_t lies in $\mathfrak{P}_{n-1}(R^1)$ for t near the end points of I_i, the induction hypothesis applies. The resulting homotopies, defined over the intervals I_i, fit together to give the desired homotopy of J_t to a path in $\mathfrak{P}_n(R^1)$.

2.8. *Proof of* 2.1. Choose b so that $f_t(x) = (x, 0)$ for $|x| > b/2$. A subdivision of the interval $[-b, b]$ induces, for each t, a polygonal approximation G_t to the curve f_t. Let the subdivision be fine enough so that the graph G_t is simple for each t, so that G_t is a differentiable path in $\mathfrak{P}_n(R^2)$. Given ε very small, let G_t^ε be the C^1 curve obtained by rounding off the corners of the graph G_t by inscribing a circle at each vertex p which is tangent to the edges incident on p at a distance ε from p. For the moment, we do not parametrize this curve.

The curve f_t may be slid along its normals onto the curve G_t^ε, at least if the subdivision defining G_t^ε is suitably fine. For each t, the parametrization of the curve G_t^ε is taken to be that induced from the curve f_t. Then G_t^ε is a loop in $\mathcal{E}^1(R^1, R^2)$. By 2.7, G_t may be deformed into a path H_t in $\mathfrak{P}_n(R^1)$; this deformation carries the curve G_t^ε onto H_t^ε for each t. Again, we take the parametrization of each curve to be that induced from G_t^ε; the result is a homotopy through differentiable loops in $\mathcal{E}^1(R^1, R^2)$ to the loop H_t^ε in $\mathcal{E}^1(R^1, R^1)$.

This does not quite prove the lemma, because the homotopy took place in $\mathcal{E}^1(R^1, R^2)$ rather than in $\mathcal{E}^2$. Let g(x, t, s) be the homotopy, with s the homotopy parameter. Then

$$h(x, t, s) = \frac{1}{2\delta} \int_{x-\delta}^{x+\delta} g(y, t, s)\, dy$$

is the desired homotopy, for δ sufficiently small.

REFERENCES

1. S. S. Cairns, *An elementary proof of the Jordan-Schoenflies theorem*, Proc. Amer. Math. Soc. 2 (1951), 860-867.

2. J. R. Munkres, *Obstructions to the smoothing of piecewise-differentiable homeomorphisms* (to appear in Ann. of Math. (2) 72 (1960)).

3. S. Smale, *Diffeomorphisms of the* 2*-sphere*, Proc. Amer. Math. Soc. 10 (1959), 621-626.

Princeton University