Critical Sets in the Plane

ALEC NORTON & CHARLES PUGH

1. Introduction

A point $x \in \mathbf{R}^m$ is *critical* for a smooth (C^∞) function $f: \mathbf{R}^m \to \mathbf{R}$ if its derivative at x is zero, $(Df)_x = 0$. If x is not critical then it is *regular*. We are interested in the topology of the set of all critical points of f,

$$\operatorname{cp}(f) := \{x \in \mathbf{R}^m : f \text{ is critical at } x\}.$$

In contrast, the (Antony) Morse–Sard theory measures the set of critical values $\operatorname{cv}(f) := f(\operatorname{cp}(f))$, while (Marston) Morse theory counts the "types" (maximum, saddle, minimum) of critical points. We say $C \subset \mathbf{R}^m$ is *critical* if $C = \operatorname{cp}(f)$ for some smooth $f: \mathbf{R}^m \to \mathbf{R}$, and *properly critical* if such an f exists which is proper, that is, inverse images of compact sets are compact. Clearly, a critical set is closed. What other properties does it have? When $m = 1$ it is easy to see that there are no other requirements—any closed set in $\mathbf{R}$ is critical. In $\mathbf{R}^2$ there is just one other requirement in the compact case.

THEOREM 1. *A compact non-empty subset of* $\mathbf{R}^2$ *is critical if and only if it is properly critical if and only if the components of its complement are multiply connected.*

Recall that a *component* of a topological space S is a maximal connected subset of S; it is *multiply connected* if it is not simply connected. Theorem 1 is proved in Section 4. For example, if C is any finite set of points or a Cantor set in the plane, then it is properly critical. Their complements are multiply connected. On the other hand, a circle is not critical. Note that the condition of multiple connectivity is a topological condition on the complement, not on C. For example, if C is the union of a circle and a point then it is critical if and only if the point is inside the circle. Similarly, for functions defined on the 2-sphere, the union of the equator and the two poles is a critical set, but not the equator and two points in the Northern hemisphere.

We permit the critical points p to be degenerate; the nondegenerate critical points (the Hessian $(D^2 f)_p$ has nonzero eigenvalues) form a discrete set, making all of our considerations trivial. Typically, the critical points we construct are infinitely flat. See Section 2.

Received May 23, 1990. Revision received December 15, 1990.
Michigan Math. J. 38 (1991).

It is for simplicity that we deal with compact sets and not closed sets. The situation with closed sets is more complicated; the generalization of Theorem 1 appears in Section 5. However, let us point out here that it is unreasonable to expect properness of f when the critical set is noncompact. For if $C=\mathrm{cp}(f)$ is closed, unbounded, and connected, then by Sard's theorem f is constant on C, say $f(C)=c$, and $f^{-1}(c)$ is noncompact, contrary to f being proper.

Proof of the easy half of Theorem 1. Suppose that U is a simply connected component of the complement of $\mathrm{cp}(f)$. Since $\mathrm{cp}(f)\neq\emptyset$, U is bounded. Let c be a regular value of f assumed on U. Then $\Gamma=f^{-1}(c)\cap U$ is a non-empty compact boundaryless 1-manifold; it consists of Jordan curves J and $f=c$ on J. At some point p inside J we have a maximum $f(p)>c$, a minimum $f(p)<c$, or $f\equiv c$ inside J. In any case $(Df)_p=0$ at some p inside J. Since U is simply connected, $p\in U$, contrary to U being in the complement of the critical set. □

The corresponding issues in dimension $m\geq 3$ are tackled in [4], but we make a few comments here. It is easy to see that any finite set in $\mathbf{R}^m$ is a critical set but that the $(m-1)$-sphere is not. Indeed, the 2-dimensional argument just given proves Theorem 2.

THEOREM 2. *If $C\subset\mathbf{R}^m$ is critical, compact, and non-empty, then any bounded component of C^c has disconnected boundary. In particular, no compact hypersurface in $\mathbf{R}^m$, smooth or not, is a critical set.*

Space curves are more fun. The function $f(x,y,z)=z(x^2+y^2-1)$ is critical exactly at the unit circle in the $z=0$ plane in $\mathbf{R}^3$. It is not proper because the whole $z=0$ plane is part of $f^{-1}(0)$. In fact, we have this next theorem.

THEOREM 3. *A circle in $\mathbf{R}^3$ is not properly critical.*

Proof. When $m\geq 2$, the point at ∞ in $\mathbf{R}^m$ has small connected neighborhoods V. Under f, the image of V is connected. By properness,

$$\inf\{|fv|: v\in V\}\to\infty \quad\text{as}\quad \inf\{|v|: v\in V\}\to\infty.$$

Since fV is connected, this implies that

$$\inf\{fv: v\in V\}\to\infty \quad\text{or}\quad \sup\{fv: v\in V\}\to-\infty.$$

Hence, the image of a proper function $f:\mathbf{R}^3\to\mathbf{R}$ is a half-ray $(-\infty,a]$ or $[a,\infty)$; we assume it is the latter and that $\mathrm{cp}(f)=C$. The minimum value a is achieved and $f^{-1}(a)$ is a non-empty subset of $\mathrm{cp}(f)$. Assume that $\mathrm{cp}(f)$ is an embedded circle C. Since C is connected, f is constant on C; that is, $f(C)=a$. Take any $b>a$ and consider $\Sigma=f^{-1}(b)$. It is a non-empty, smooth, boundaryless, compact surface in $\mathbf{R}^3$. Inside each component of Σ, f achieves a maximum or minimum; that is, inside each component of Σ there lies a point of C. Since C is connected and disjoint from Σ, we see that Σ is connected and encloses C.

Properness of f implies that the gradient flow (i.e., the flow generated by the vector field $\mathrm{grad}(f)$) has the property that each of its trajectories off C meets Σ. It follows that C^c is diffeomorphic to $\Sigma \times \mathbf{R}$. Adjoining the point at infinity to $\mathbf{R}^3$ produces the 3-sphere. The fact that $C^c \approx \Sigma \times \mathbf{R}$ implies that the neighborhood of ∞ is foliated by copies of Σ. This can only happen if $\Sigma \approx S^2$. However, $S^2 \times \mathbf{R}$ is simply connected, whereas $\mathbf{R}^3$ minus an embedded circle is never simply connected. Having arrived at a contradiction, the theorem is proved. □

Theorem 3 shows two ways in which Theorem 1 fails in higher dimensions. First, proper criticality and criticality for compact sets in $\mathbf{R}^m$ when $m \geq 3$ are inequivalent, since the circle is critical but not properly so. Second, multiple connectedness of the complement of a compact set does not imply that it is critical. For we could take the circle C and enclose it in a large sphere S, and then draw a circle C' outside S. The set $K = C \cup S \cup C'$ is compact, and both components of its complement are multiply connected in the sense that their homology and homotopy groups H_1, H_2, π_1, π_2 are nontrivial. Suppose that $f: \mathbf{R}^3 \to \mathbf{R}$ is a smooth function with $\mathrm{cp}(f) = K$. Because C and S are connected, f is constant on each; say $f(C) = c$ and $f(S) = 0$. Let B be the compact ball bounded by S. Since f is nonconstant on B, we may assume that $\max f|_B > 0$. The maximum occurs at some point p interior to B and therefore on C. Thus, $c > 0$. Similarly, $0 < f(x) < c$ for all $x \in B \backslash K$. Thus, for $x \in B$, $f(x) \to 0$ if and only if $x \to S$. The interior of B is diffeomorphic to $\mathbf{R}^3$ by a diffeomorphism, say $\varphi: \mathrm{int}(B) \to \mathbf{R}^3$, fixing C. Then $F(x) = \log(f \circ \varphi^{-1}(x))$ is a smooth proper function on $\mathbf{R}^3$ with critical set C, and this contradicts Theorem 3.

A more general version of Theorem 3 is the following.

THEOREM 4. *A non-empty compact connected set $C \subset \mathbf{R}^m$ is properly critical if and only if it is* smoothly cellular, *that is, its complement is diffeomorphic to the complement of a point.*

Proof. The "only if" part of the theorem has a proof similar to that of Theorem 3. Let $C = \mathrm{cp}(f) = f^{-1}(a)$ for $f: \mathbf{R}^m \to [a, \infty)$ a smooth proper function, and consider $\Sigma = f^{-1}(b)$ as before. Then $\mathbf{R}^m \backslash C \approx \Sigma \times \mathbf{R}$, where $\Sigma \times \infty$ corresponds to $\infty \in S^m$. It is tempting to conclude that "clearly" $\Sigma \approx S^{m-1}$ and hence that $\mathbf{R}^m \backslash C \approx S^{m-1} \times \mathbf{R} \approx \mathbf{R}^m \backslash 0$. Here is a more convincing argument.

Choose a smooth compact disc $D \subset S^m$ such that $\infty \in \mathrm{int}(D)$ and $D \cap C = \emptyset$. (See Figure 1.) Every trajectory $\varphi_t x$ of the $\mathrm{grad}(f)$-flow in $\mathbf{R}^m \backslash C$ crosses ∂D on its way from the source C to the sink ∞. Thus, the closure of the reverse orbit of ∂D, $\vartheta_-(\partial D) = \{\varphi_t x : t \leq 0 \text{ and } x \in \partial D\}$, is a compact subset of $\mathbf{R}^m$. Let $\beta: S^m \to [0, 1]$ be a smooth bump function which is 1 on some neighborhood of $\vartheta_-(\partial D) \cup C$ and 0 on some neighborhood W of ∞. Then βf is smooth on S^m (even though f may be nasty at ∞) and $-\mathrm{grad}(\beta f)$ generates a smooth flow ψ on S^m. It satisfies

(a) $\psi_t(x) = \varphi_{-t}(x)$ for $x \in \vartheta_-(\partial D)$ and $t \geq 0$, and

(b) $\psi_t(x) = x$ for $x \in C \cup W$ and $t \in \mathbf{R}$.

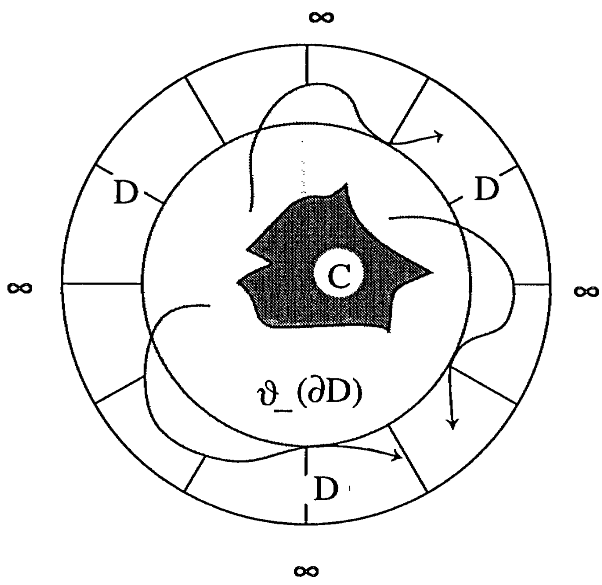

Figure 1

If T is large then $\Psi = \psi_T : S^m \to S^m$ is a diffeomorphism such that $\Psi(x) = x$ on $C \cup W$ and $\{\Psi^n(D)\}$ is a monotone family of smooth discs with $\bigcup \Psi^n(D) = S^m \backslash C$. (See Figure 2.) In particular, $S^m \backslash C$ is the monotone union of the smooth open m-cells $\Psi^n(\operatorname{int}(D))$. By a C^∞ version of M. Brown's monotone cell theorem (proved in the appendix), $S^m \backslash C$ is diffeomorphic to $\mathbf{R}^m$ and therefore $\mathbf{R}^m \backslash C \approx \mathbf{R}^m \backslash 0$.

To prove the converse, assume that $\varphi : \mathbf{R}^m \backslash C \to \mathbf{R}^m \backslash 0$ is a diffeomorphism. Reflecting φ in the unit sphere if necessary, we may assume that φ extends to a continuous map $\bar{\varphi} : \mathbf{R}^m \to \mathbf{R}^m$ sending C to 0. Choose a smooth proper function $g : \mathbf{R}^m \to \mathbf{R}$ having a very flat minimum at 0 and having no other critical points. Then $f = g \circ \bar{\varphi}$ is smooth and proper, and $\operatorname{cp}(f) = C$.

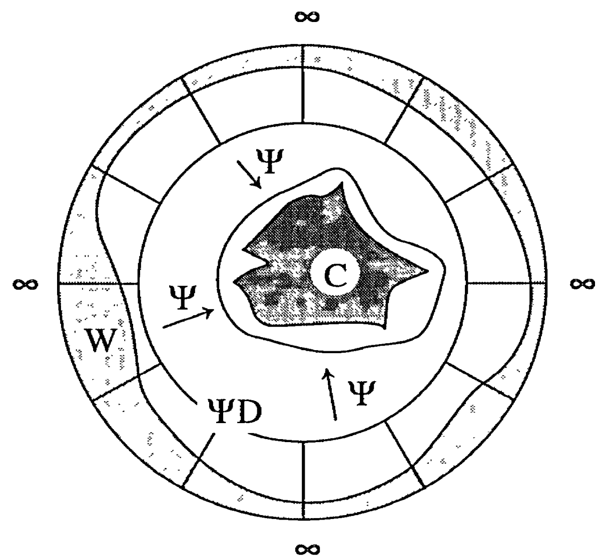

Figure 2

That the minimum of g being flat enough implies smoothness of f appears as Lemma 3 in the next section. □

We thank M. Brown, R. Gompf, and M. Shub for useful conversations.

2. Flat Critical Points

Let $g: \mathbf{R}^m \to \mathbf{R}^n$ be any function, continuous or not, and let $A \subset \mathbf{R}^m$ be a closed set. We say that g is *k-flat* at A if

$$\delta_k(x,y) := \frac{|gx-gy|}{|x-y|^k} \to 0 \quad \text{as} \quad \inf_{a \in A}\{|x-a|+|y-a|\} \to 0.$$

In other words, as (x,y) approaches the A-diagonal $\Delta_A = \{(a,a) \in \mathbf{R}^m \times \mathbf{R}^m\}$, the variation of g over the segment $[x,y]$ becomes much less than its kth power. We say that g is *∞-flat* at A if it is k-flat for all $k \geq 0$. More generally, recall that a *modulus of continuity* is a function $\mu(s) > 0$, defined for $s > 0$, such that $\mu(s) \to 0$ as $s \to 0$. The function g is *μ-flat* at A if

$$\delta_\mu(x,y) := \frac{|g(x)-g(y)|}{\mu(|x-y|)} \to 0 \quad \text{as} \quad (x,y) \to \Delta_A.$$

If g is C^1 then g is critical at p if and only if it is 1-flat at p. If $\mathbf{R}^m$ and $\mathbf{R}^n$ are replaced by smooth manifolds M^m and N^n, then we make the corresponding definition using smooth compact charts. It is easy to see that μ-flatness is independent of the particular coverings considered.

The idea of flatness and its use to smooth off corners as in Lemma 3 below are reminiscent of the Kneser–Glaeser rough composition theorem. See [1, p. 35]. However, we are concerned with the differentiability of flat compositions, not their existence, and this makes our results more elementary.

First we prove an existence result for uncomposed flat functions.

LEMMA 1. *Given a compact set $K \subset \mathbf{R}^m$ and a modulus of continuity μ, there is a smooth proper function $g: \mathbf{R}^m \to [0,\infty)$ such that $g^{-1}(0) = K$ and g is μ-flat at K. If $m=1$ then $f = \int g$ has* $\mathrm{cp}(f) = K$ *and is μ-flat at K.*

Proof. This is an easy differential topology exercise. Choose a locally finite partition of unity, $1 = \sum \sigma_i$ on $\mathbf{R}^m \setminus K$, such that the support of each σ_i is disjoint from K. Define $g = \sum \epsilon_i \sigma_i$, where $\epsilon_i \to 0$ rapidly as $\mathrm{supp}(\sigma_i) \to K$ and $\epsilon_i \to \infty$ as $\mathrm{supp}(\sigma_i) \to \infty$. The assertion in dimension 1 is clear (it includes Theorem 1 on $\mathbf{R}$). □

LEMMA 2. *If $g: \mathbf{R}^m \to \mathbf{R}^n$ is ∞-flat at a closed set $A \subset \mathbf{R}^m$ and if g is smooth on A^c, then g is smooth on all of $\mathbf{R}^m$ and all its derivatives vanish on A.*

Proof. ∞-flatness implies 1-flatness, so if $a \in A$ and $v \in \mathbf{R}^m$ then

$$\frac{|g(a+tv)-g(a)|}{|t|} = \delta_1(a+tv,a)|v| \to 0$$

as $t \to 0$. Convergence is uniform in v, $|v| \le 1$. This implies that $(Dg)_a$ exists and equals zero. Thus, $Dg: \mathbf{R}^m \to L(\mathbf{R}^m, \mathbf{R}^n)$ is a well-defined map that is smooth on A^c. We claim that Dg is ∞-flat at A.

For $x, v \in \mathbf{R}^m$, we have

$$(Dg)_x(v) = \frac{g(x+tv)-g(x)}{t} + r(x,v,t)$$

and $r/t \to 0$ as $t \to 0$. Convergence is uniform in v, $|v| \le 1$, but it conceivably depends on x. If $(x,y) \to \Delta_A$ with $x, y \in A^c$ and $x \ne y$, choose $t = t(x,y)$ such that

$$0 < t \le |x-y|, \quad \frac{r(x,v,t)}{t} \le |x-y|^{k+1}, \quad \text{and} \quad \frac{r(y,v,t)}{t} \le |x-y|^{k+1}.$$

Then

$$\begin{aligned}
&\frac{|(Dg)_x(v)-(Dg)_y(v)|}{|x-y|^k} \\
&\qquad = \frac{\left| \frac{g(x+tv)-g(x)}{t} + r(x,v,t) - \frac{g(y+tv)-g(y)}{t} - r(y,v,t) \right|}{|x-y|^k} \\
&\qquad \le \delta_{k+1}(x+tv,x)|v|^{k+1} + \delta_{k+1}(y+tv,y)|v|^{k+1} + 2|x-y|,
\end{aligned}$$

and since $0 < t \le |x-y|$, this quantity tends to zero as $(x,y) \to \Delta_A$. If x or y belongs to A then $(Dg)_x$ or $(Dg)_y$ equals zero, and the preceding estimate improves. Thus Dg is k-flat at A for all k, and we can apply induction to conclude that $D(Dg)$ exists everywhere and vanishes on A. Continuing, we see that g is smooth and that $(D^k g)_a = 0$ for all $a \in A$ and $k \ge 1$. □

We shall prove that pulled-back and pushed-ahead functions $\varphi^* g = g \circ \varphi$ and $\varphi_* g = g \circ \varphi^{-1}$ are smooth when φ is not entirely smooth but g is very flat.

LEMMA 3. *Suppose that A, B are compact subsets of smooth manifolds M^m, N^n, and $\varphi: M \to N$ is a continuous map sending A into B, which is smooth off A. There exists a modulus of continuity ν such that if $g: N \to \mathbf{R}$ is smooth and ν-flat at B then the pulled-back function $\varphi^* g = g \circ \varphi: M \to \mathbf{R}$ is smooth. If in addition φ diffeomorphs A^c onto B^c then, for some modulus of continuity μ, the pushed-ahead function $\varphi_* g = g \circ \varphi^{-1}: N \to \mathbf{R}$ is smooth, provided that $g: M \to \mathbf{R}$ is smooth, constant on A, and μ-flat at A.*

Proof. The pulled-back function $\varphi^* g = g \circ \varphi$ is intrinsically defined on M. Since smoothness and flatness are local issues, we may assume that M, N are smooth compact manifolds with boundary, that they are contained in $\mathbf{R}^m$, $\mathbf{R}^n$, and that they contain A, B in their interiors. Working in a single coordinate system like this is mainly a notational simplification. Define

$$\nu(s) = \inf\{e^{-1/|x-y|} : x, y \in M \text{ and } |\varphi(x) - \varphi(y)| \ge s\}.$$

Then $\nu(|\varphi x - \varphi y|) \le e^{-1/|x-y|}$. If g is ν-flat at B, then $g \circ \varphi$ is ∞-flat at A and (by Lemma 2) smooth on M. That is, $\varphi^* g$ is smooth.

The analysis of the push-ahead $\varphi_* g = g \circ \varphi^{-1}: N \to \mathbf{R}$ is similar. Again, making everything local, we assume that M, N are compact smooth m-manifolds with boundary, contained in $\mathbf{R}^m$, and containing A, B in their interiors. Because g is constant on A, $\varphi_* g$ is well defined. We set

$$\mu(s) = \inf\{e^{-1/|x-y|}: x, y \in N \text{ and } |\varphi^{-1}(x) - \varphi^{-1}(y)| \geq s\}.$$

Then $\mu(|\varphi^{-1}(x) - \varphi^{-1}(y)|) \leq e^{-1/|x-y|}$. If $g: M \to \mathbf{R}$ is μ-flat at A then $g \circ \varphi^{-1}$ is ∞-flat at B and, by Lemma 2, $g \circ \varphi^{-1}$ is smooth. □

REMARK. Observe that $\varphi^* g$ and $\varphi_* g$ are as flat as we want at A and B by proper choices of ν and μ.

3. Some Plane Topology

In this section we do some standard 2-dimensional pushing and pulling. Any collection $\mathcal{J}$ of disjoint Jordan curves in $\mathbf{R}^2$ has a natural partial order. One writes $J < J'$ to indicate that J is *enclosed* by J'. If neither $J < J''$ nor $J'' < J$ then one says that J, J'' are *mutually exterior.* (See Figure 3.) Clearly, distinct maximal elements of $\mathcal{J}$ respecting the partial order $<$ are mutually exterior.

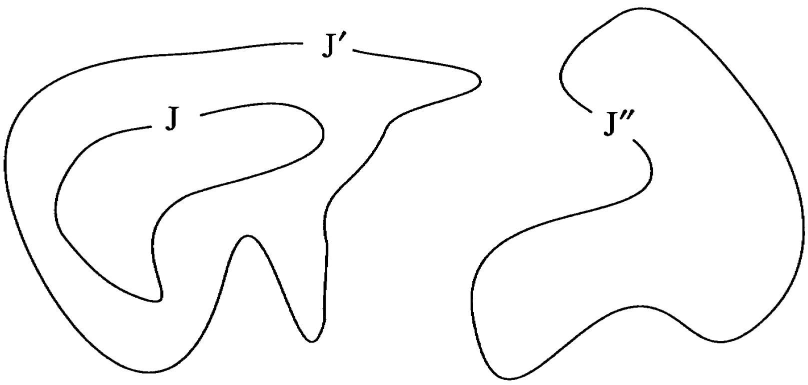

Figure 3

LEMMA 4. (See Figure 4.) *Let $\mathcal{J}$ be a finite collection of disjoint Jordan curves in $\mathbf{R}^2$. The subcollection of maximal elements respecting the partial order $<$ forms the frontier of the unbounded component of $\mathbf{R}^2 \setminus \bigcup_{J \in \mathcal{J}} J$.*

Proof. Clear. □

LEMMA 5. *Given a compact set $K \subset \mathbf{R}^2$ and given $\epsilon > 0$, there exist finitely many mutually exterior smooth Jordan curves $J_1, \ldots, J_k$ such that:*

(a) *$J_1, \ldots, J_k$ are disjoint from K but are contained in the ϵ-neighborhood of K; and*

(b) *$K \subset \bigcup D_i$, where $\partial D_i = J_i$.*

Proof. Let $g: \mathbf{R}^2 \to [0, \infty)$ be the function supplied by Lemma 1. It is smooth and proper, and $g^{-1}(0) = K$. By the Morse–Sard theorem, most of its values

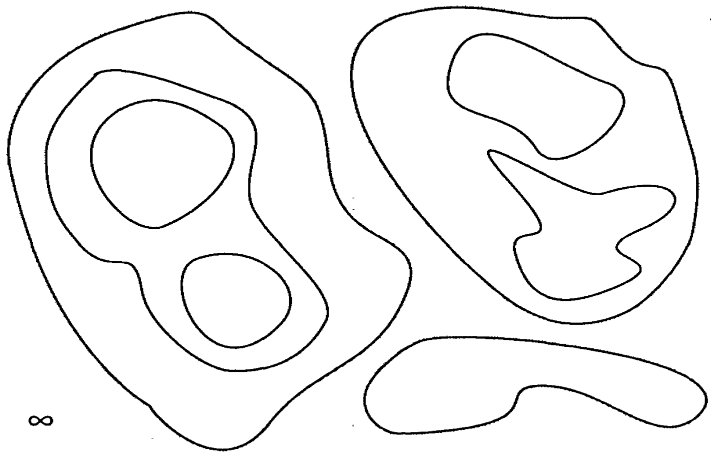

Figure 4

are regular. Let $c>0$ be a small regular value. Then $\Gamma = g^{-1}(c)$ is a compact smooth 1-manifold that lies in the ϵ-neighborhood of K. Necessarily, Γ is disjoint from K, separates K from ∞, and consists of finitely many smooth Jordan curves $\gamma_1, \dots, \gamma_m$. By Lemma 4, there are finitely many mutually exterior Jordan curves $J_1, \dots, J_k$ among $\gamma_1, \dots, \gamma_m$ such that $\bigcup J_i$ separates K from ∞. Thus K is contained in the discs D_i bounded by the J_i. Since the J_i are part of Γ, they lie in the ϵ-neighborhood of K and miss K. □

LEMMA 6 (Punctures). *Any non-empty connected open subset U of S^2 is diffeomorphic to a punctured 2-sphere where the punctures form a totally disconnected compact subset K of a longitude. Besides, a diffeomorphism $\varphi: U \to S^2 \setminus K$ can be chosen so that it extends to a continuous map $\bar{\varphi}: S^2 \to S^2$, isotopic to the identity, which pinches distinct connected components of $S^2 \setminus U$ to distinct points of K.*

REMARK. The set U can be quite nasty—it could be a disc minus $C \times [0,1]$ where C is a Cantor set, or it could be a disc with a dendrite or non-locally connected boundary minus a set of holes that accumulate densely at the boundary. (See Figure 5.)

Proof. Let P be the complement of U; that is, let $P = \mathbf{R}^2 \setminus U$. We may assume that P is interior to the unit disc $\mathbf{D}$. Choose $\epsilon_0 > 0$ with $\epsilon_0 < d_{\min}(P, \partial\mathbf{D})$, where $d_{\min}$ denotes the minimum distance between points of P and $\partial\mathbf{D}$. By Lemma 5, there exist disjoint discs $D_1, \dots, D_k$ whose union contains P, and ∂D_i is a smooth Jordan curve lying in the ϵ_0-neighborhood of K with $\partial D_i \cap P = \emptyset$.

Set $D^0 = \bigcup D_i$; it is a compact smooth neighborhood of P. Then choose $\epsilon_1 > 0$ with $\epsilon_1 < d_{\min}(P, \partial D^0)$. Inside each D_i, we use Lemma 5 on $D_i \cap P$ to find a finer family of discs covering P; this time their boundaries lie in the ϵ_1-neighborhood of P. Since $\partial D_i \cap P = \emptyset$, this is valid. Repeating the construction, we get a sequence $\{\mathfrak{D}^m\}$ of finite covers of P, $\mathfrak{D}^m = \{D_1^m, \dots, D_k^m\}$, where D_i^m is a smooth disc and $k = k(m)$. We may assume each D_i^m meets P.

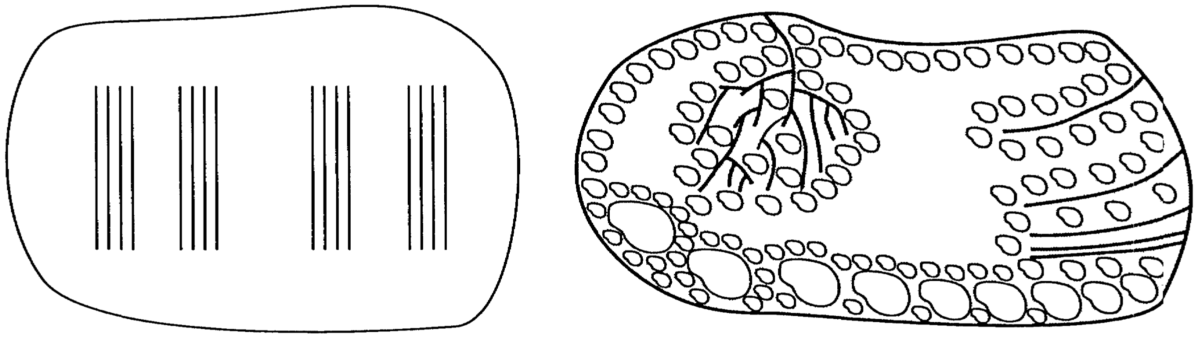

Figure 5

Set $D^m = \bigcup_i D_i^m$. The D^m are a nested strictly decreasing sequence of compact smooth neighborhoods of P such that ∂D^m is contained in the ϵ_m-neighborhood of P. We may assume that $\epsilon_m \to 0$ as $m \to \infty$.

For each $m = 0, 1, 2, \ldots,$ choose an index $i = i(m) \in \{1, \ldots, k(m)\}$ and set

$$Q_m = D_{i(m)}^m \quad \text{and} \quad Q = \bigcap Q_m.$$

We claim that if $Q \neq \emptyset$ then it is a component of P. Clearly, Q_m is compact, connected, and splits P; that is, each component of P lies wholly in Q_m or wholly in its complement.

If Q is not a single component of P then there exists a point $x \in Q \backslash P$. By hypothesis, $P^c = U$ is a connected neighborhood of ∞, and so there is a path γ in U from x to ∞. For some $\epsilon > 0$, γ misses the ϵ-neighborhood of P, and so $\gamma \cap \partial D^m = \emptyset$ for large m. In particular, $\gamma \cap \partial Q_m = \emptyset$. But since $x \in Q$, $x \in Q_{m+1}$. Since Q_{m+1} is a disc that is interior to Q_m, there is a path β in Q_{m+1} from $P \cap Q$ to x. (See Figure 6.) Then $\beta \cup \gamma$ is a path from P to ∞ that misses ∂D^m, and contradicts the fact that $P \subset D^m$. This proves the claim that Q is a component of P.

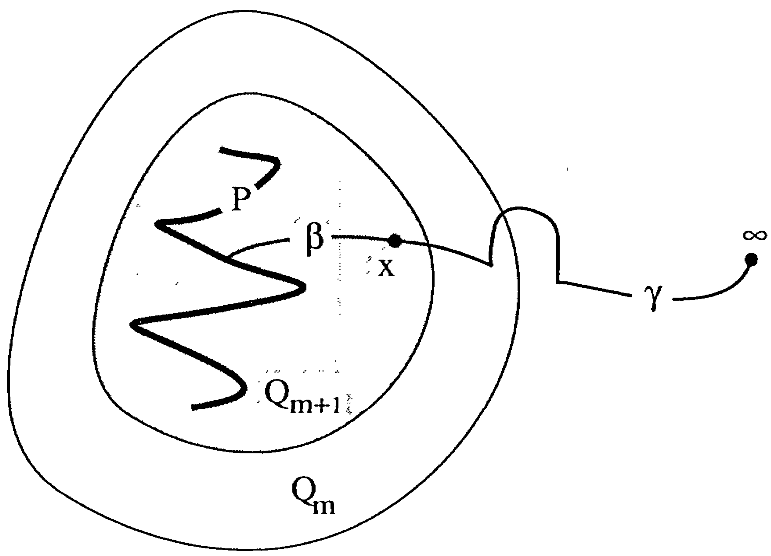

Figure 6

To complete the proof of Lemma 6, we must construct a diffeomorphism of U to $S^2\backslash K$, where K is some totally disconnected subset of a longitude L. We do so by stages. Let ψ_0 be a diffeomorphism of S^2 to itself, isotopic to the identity, that sends each disc D_i^0 of D^0 to a disc Δ_i^0 that is small and has nontrivial intersection with L. Then let ψ_1 be a diffeomorphism of S^2 to itself that sends each disc Δ_i^0 to itself and sends each disc $\psi_0(D_j^1)$ to a much smaller disc Δ_j^1 also having nontrivial intersection with L. (See Figure 7.) Off $\Delta^0 = \bigcup_i \Delta_i^0$, let ψ_1 be the identity map. (We are using the fact that there is no obstruction to moving several disjoint subdiscs to new positions in a larger disc.)

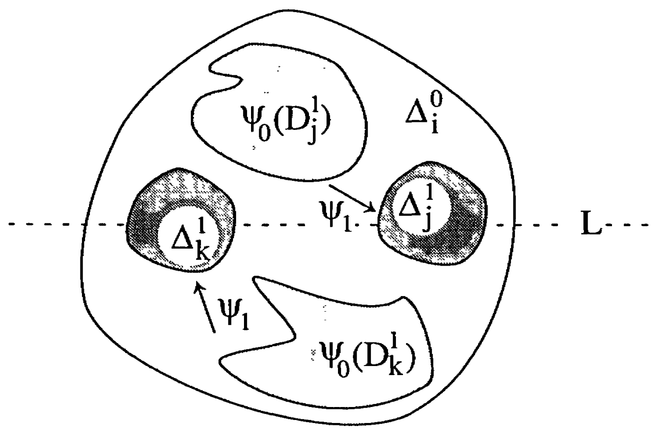

Figure 7

Inductively assume that $\psi_0, \ldots, \psi_m$ have been defined and set

$$\Psi_m = \psi_m \circ \cdots \circ \psi_0.$$

Let ψ_{m+1} be a diffeomorphism of S^2 to itself that sends each disc Δ_i^m to itself and sends the subdisc $\Psi_m(D_j^{m+1})$ to a much smaller subdisc Δ_j^{m+1} having nontrivial intersection with L. We may ensure that $\operatorname{diam}(\Delta_i^{m+1}) < 1/(m+1)$. Off $\Delta^m = \bigcup_i \Delta_i^m$, make ψ_{m+1} the identity map.

The composition diffeomorphisms Ψ_m and Ψ_{m+1} differ only on D^m, and the difference of their values on D_i^m is confined to Δ_i^m. Since $\operatorname{diam}(\Delta_i^m) \rightrightarrows 0$ as $m \to \infty$, it follows that Ψ_m converges uniformly to a limit as $m \to \infty$, say $\Psi = \lim \Psi_m$, and Ψ diffeomorphs $U = \bigcup_m (D^m)^c$ onto $V = \bigcup_m (\Delta^m)^c$. By construction, $K = \bigcap_m \Delta^m$ is a compact totally disconnected subset of L and $K = V^c$. Because infinite nested intersections of the discs D_i^m are components of P, and infinite nested intersections of the discs Δ_i^m are points of K, we see that Ψ sends distinct components of P to distinct points of K. The diffeomorphisms $\psi_1, \psi_2, \ldots$ are isotopic to the identity because they equal the identity on parts of S^2. The diffeomorphism ψ_0 is isotopic to the identity by choice. Thus, $\Psi = \lim \psi_m \circ \cdots \circ \psi_0$ is isotopic to the identity. □

4. Compact Planar Critical Sets

In this section we prove the harder half of Theorem 1: *A compact subset of the plane is critical if each component of its complement is multiply connected.* If $f: M \to [0,1]$ is smooth and $\operatorname{cp}(f) = U^c$, then we say that f is a *boundary-critical Morse function* for U. Inside U, f has no critical points, while outside U, all points are critical. On connected components of U^c, f is constant. The *exterior* of a bounded set $B \subset \mathbf{R}^m$ is defined to be the connected component of ∞ in B^c; we denote it by $\mathrm{ext}(B)$.

LEMMA 7. *Open multiply connected subsets of S^2 have boundary-critical Morse functions.*

Proof. First assume that the open, multiply connected set $U = \Sigma \setminus K$, where K is a relatively closed, totally disconnected subset of the interval $(-1,1) \times \{0\}$ in the x-axis and Σ is the open square $\Sigma = (-1,1) \times (-1,1)$. Since U is multiply connected, $K \neq \emptyset$; we assume that K contains the origin. By Lemma 1 there is a smooth function $g: \mathbf{R} \to [0,1]$ such that

$$g^{-1}(0) = K^* = \{x \in \mathbf{R}: (x,0) \in K \text{ or } |x| \geq 1\},$$

and g is very flat at K^*. Define

$$G_-(x) = a \int_{-1}^{x} g(t)\,dt \quad \text{and} \quad G_+(x) = \int_x^1 g(t)\,dt,$$

where a is a constant chosen so that $G_-(0) = G_+(0)$. We glue the functions $G_\pm$ together as

$$G(x) = \begin{cases} G_-(x) & \text{if } x \leq 0, \\ G_+(x) & \text{if } x \geq 0. \end{cases}$$

Then $G: \mathbf{R} \to [0,1]$ is smooth, positive on $(-1,1)$, and zero elsewhere; G is very flat at K^*; and $\mathrm{cp}(G) = K^*$. Let $\beta: \mathbf{R} \to [0,1]$ be a smooth bump function such that β is positive on $(-1,1)$, zero on the rest of $\mathbf{R}$, and has a unique critical point in $(-1,1)$ at 0. We may make this critical maximum as flat as we want. Let $B = \mathrm{cp}(\beta) = \{y: |y| \geq 1 \text{ or } y = 0\}$. We then set

$$h(x,y) = G(x)\beta(y).$$

It is clear that $h: \mathbf{R}^2 \to [0,1]$ is a smooth function, positive on Σ and zero elsewhere. If $(x,y) \in \Sigma$ is a critical point of h then both partials of h vanish there. This happens if and only if $y = 0$ and $(x,0) \in K$. That is, $\mathrm{cp}(h) = K \cup \mathrm{ext}(\Sigma)$. This takes care of the case $U = \Sigma \setminus K$. Note that since G and β are very flat at K^* and B, respectively, h is very flat at the critical set $K \cup \mathrm{ext}(\Sigma)$.

Now consider the general case where U is a multiply connected open subset of S^2. By Lemma 6 there is a continuous map $\varphi: S^2 \to S^2$ such that U is carried diffeomorphically to the complement of a compact totally disconnected subset $K \subset L$, where L is some longitude of S^2. The components of U^c are carried to distinct points of K. There is also a map $\psi: S^2 \to S^2$ that diffeomorphs Σ onto $S^2 \setminus \infty$, sends the exterior of Σ to ∞, and sends the x-axis to the longitude L. We consider the subset $K' = \psi^{-1}(K)$ of the x-axis, and apply the previous case to obtain a smooth function $h: \mathbf{R}^2 \to [0,1]$ such that h is very flat at $K' \cup \mathrm{ext}(\Sigma)$, $h^{-1}(0) = \mathrm{ext}(\Sigma)$, and $\mathrm{cp}(h) = K' \cup \mathrm{ext}(\Sigma)$.

First we apply Lemma 3 and the subsequent remark with $M = N = S^2$, $A = \mathrm{ext}(\Sigma)$, and $B = \{\infty\}$. Since h is constant on A and h is very flat at A, we know that h pushes ahead to a smooth function, $\psi_* h\colon S^2 \to [0,1]$, where $\psi_* h$ is very flat at ∞ and $\mathrm{cp}(\psi_* h) = K$. Then we apply Lemma 3 with $M = N = S^2$, $A = U^c$, and $B = K$. We see that $\psi_* h$ pulls back to a smooth function $\varphi^*(\psi_* h) = f\colon S^2 \to [0,1]$, and that f is very flat at U^c. In particular, $\mathrm{cp}(f) \supset U^c$. On the other hand, each $x \in U$ is carried by φ to a point of $S^2 \backslash K$ and then by ψ^{-1} to a point of $\Sigma \backslash K'$. (See Figure 8.) It follows that x is regular and $\mathrm{cp}(f) = U^c$. □

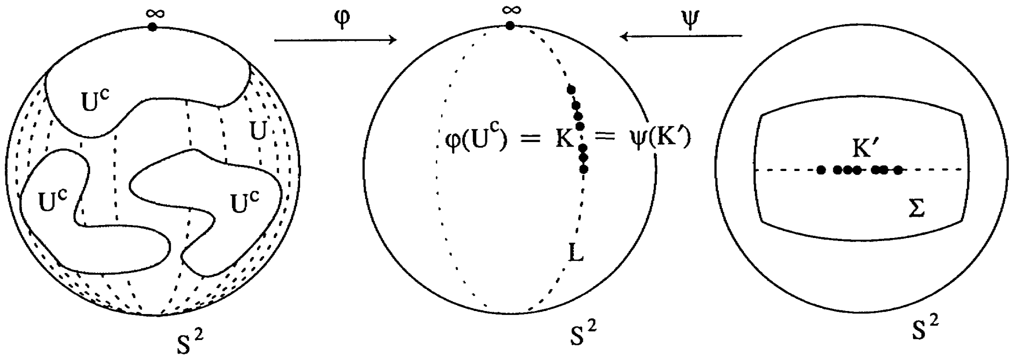

Figure 8

EXAMPLE. If U is a 2-holed annulus (a "pair of pants"; see Figure 9), then Lemma 7 produces a function critical only on U^c. Its graph is a double volcano with no saddle between the craters.

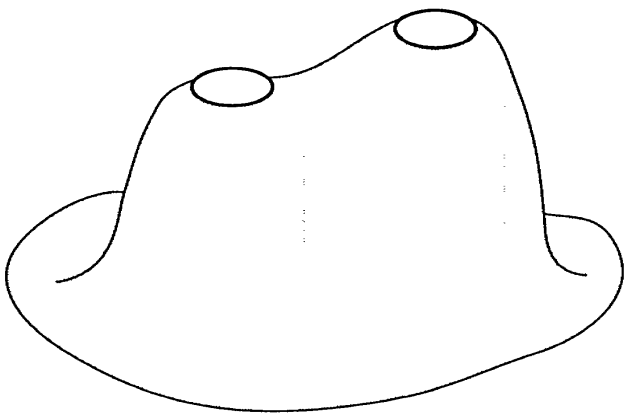

Figure 9

Proof of Theorem 1. Let K be a compact non-empty subset of the plane. In Section 1 we showed that if K is critical then each component of K^c is multiply connected. (This was easy.) Now we prove the converse.

We consider the components of K^c, and label them $U_0, U_1, \ldots$, where $U_0 = \mathrm{ext}(K)$ is the unique unbounded component of K^c. Each U_j is multiply connected and we can use Lemma 7 to find an appropriate boundary-critical

Morse function f_j for it. On U_j, f_j has no critical points, but f_j is constant on each component of U_j^c. Now any other U_k lies entirely in one component of U_j^c because it is connected. Thus, f_j is constant on each U_k with $k \neq j$.

For $j=0$, we may assume that $f_0^{-1}(0)=\infty$. For $j \geq 1$, choose constants $a_j > 0$ such that the C^j size of $a_j f_j$ is $\leq 2^{-j}$. The series

$$f(x)=\log f_0(x)+\sum a_j f_j$$

converges uniformly in the C^∞ sense. It can be differentiated termwise, and we see that

$$(Df)_x=\begin{cases} [1/f_0(x)](Df_0)_x & \text{if } x \in U_0, \\ a_j(Df_j)_x & \text{if } x \in U_j \text{ and } j \geq 1, \\ 0 & \text{if } x \in K. \end{cases}$$

Thus, $\operatorname{cp}(f)=K$. Since $f_0(x) \to 0$ if and only if $|x| \to \infty$, $\log f_0$ is proper and so is f. □

5. Unbounded Planar Critical Sets

Given a closed, noncompact $K \subset \mathbf{R}^2$, when is K critical? That is, when is there a smooth function $f: \mathbf{R}^2 \to \mathbf{R}$ such that $K=\operatorname{cp}(f)$? We say that ∞ is *arcwise accessible* in $U \subset \mathbf{R}^m$ if there is an arc $\alpha: [0, \infty) \to U$ such that $\alpha(t) \to \infty$ as $t \to \infty$.

THEOREM 5. *A closed set $K \subset \mathbf{R}^2$ is critical if and only if ∞ is arcwise accessible in each simply connected component of K^c.*

An example of an unbounded connected open set in $\mathbf{R}^2$ in which ∞ is not arcwise accessible is given by

$$W=\left\{(x,y) \in \mathbf{R}^2 : 0 < y-\frac{1}{x}\sin\frac{1}{x} < 1 \text{ and } 0 < x < 1\right\}.$$

(See Figure 10.) Topologically, W is an open disc, and clearly ∞ is not arcwise accessible in W. By Theorem 5, W^c cannot be critical. One may also check this directly: If $\operatorname{cp}(f) \supset W^c$ then f is constant on ∂W, and (some thought shows that) an interior maximum or minimum is forced on it.

If K is any closed set of parallel lines in $\mathbf{R}^2$ then the components of its complement are open strips. They are simply connected, and it is easy to see directly that any such K is a critical set. This also follows from Theorem 5 since ∞ is arcwise accessible in the strips.

REMARK. At one time we believed in a different way to characterize the simply connected components U of the complement to a critical set. We thought they were "properly unbounded", by which we meant that, for any compact set B, $U \backslash B$ has an unbounded component. Although it is trivial that arcwise accessibility of ∞ implies proper unboundedness, the converse

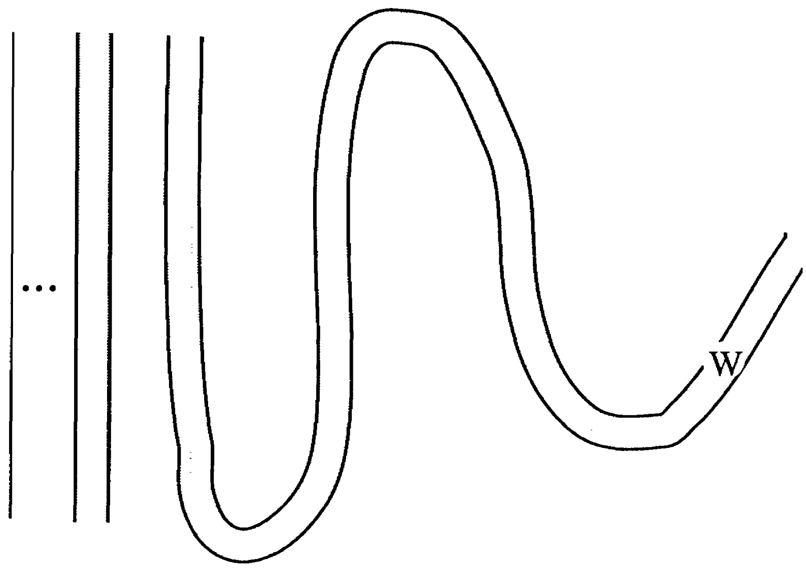

Figure 10

is false. An example can be constructed by gluing countably many elongated copies of W to a common disc.

LEMMA 8. *There exists a boundary-critical Morse function $f: \mathbf{R}^2 \to [0,1]$ for any open connected subset $U \subset \mathbf{R}^2$ in which ∞ is arcwise accessible.*

Proof. If U is multiply connected then Lemma 7 applies and produces a boundary-critical Morse function defined on all of S^2, not just on $\mathbf{R}^2$. Suppose that U is simply connected and that α is an arc to ∞ in U, $\alpha: [0, \infty) \to U$. Then $U \backslash \{\alpha(0)\}$ is multiply connected and it has a boundary-critical Morse function h defined on S^2. Excising α from U leaves a simply connected set U'; indeed, there is a continuous map $\varphi: S^2 \to S^2$ which retracts α to ∞. (See Figure 11.) It can be arranged that φ is smooth except at ∞, is the identity map off a thin neighborhood of α, and diffeomorphs $\alpha^c = \mathbf{R}^2 \backslash \alpha$ to $\mathbf{R}^2$. In particular, it diffeomorphs U' onto U. Since the restriction of φ to α^c is a diffeomorphism onto $\mathbf{R}^2$, we see that the push-ahead of $h|_{\alpha^c}$ under φ is the boundary-critical Morse function $f: \mathbf{R}^2 \to [0,1]$ that we seek. □

Proof of Theorem 5. First, suppose that U is a simply connected component of the complement of $\mathrm{cp}(f)$, where $f: \mathbf{R}^2 \to \mathbf{R}$ is smooth. Choose a regular value c assumed by f on U. Then $\Gamma = f^{-1}(c) \cap U$ is a smooth, nonempty, boundaryless 1-manifold. It is properly embedded since c is regular. This means that Γ does not accumulate on itself and is a closed subset of $\mathbf{R}^2$. Necessarily, Γ consists of Jordan curves and arcs that connect ∞ to itself. But as shown in Section 1, a Jordan curve $J \subset \Gamma$ is impossible since f would have to have a maximum or minimum inside J, say at p, and p would lie in U since U is simply connected. Therefore Γ consists of arcs to ∞ in U, demonstrating that ∞ is arcwise accessible in U.

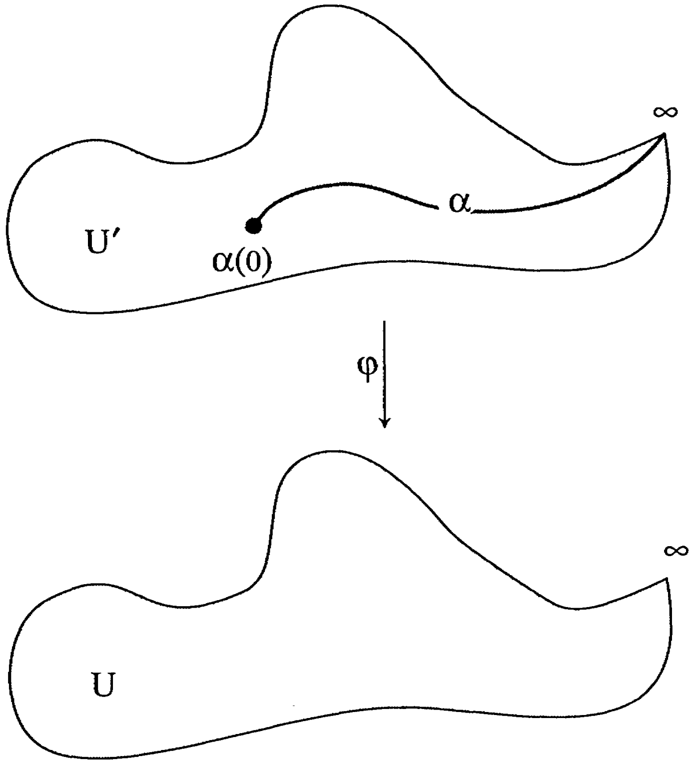

Figure 11

Second, suppose that K is a closed subset of $\mathbf{R}^2$ and ∞ is arcwise accessible in each simply connected component of K^c. Let $U_1, U_2, \ldots$ denote the components of K^c. By Lemmas 7 and 8 there exists a boundary-critical Morse function $f_j: \mathbf{R}^2 \to [0,1]$ for each U_j. Choose a constant $a_j > 0$ so that the C^j-size of $a_j f_j$, restricted to the disc of radius j, is less than 2^{-j}. Then $\sum f_j$ converges C^∞-uniformly on compact subsets of $\mathbf{R}^2$, and termwise differentiation is legal. Since f_j is constant on any U_k, $k \neq j$, we see that $\operatorname{cp}(f) = K$ as in the compact case. □

FINAL REMARK. One might ask what happens if we require that f be real analytic. Clearly, the critical set is an analytic curve or variety (with singularities permitted). One could conjecture that any analytic curve in the plane whose simply connected complementary components are arcwise accessible to ∞ is critical for some real analytic function on $\mathbf{R}^2$. Similarly, one could ask about such global conditions on the critical set of a harmonic function. However, this is only interesting in dimension at least 3, since the critical point set of a nonconstant harmonic function on $\mathbf{R}^2$ is discrete.

Appendix: Cellularity

In [3] Brown introduced the concept of (topological) *cellularity* of a compact subset $C \subset \mathbf{R}^m$. One may consider six additional types of cellularity:

(1) Brown's definition: C is the monotone intersection of compact topological m-cells.

(2) $\mathbf{R}^m \setminus C \cong \mathbf{R}^m \setminus 0$, where "$\cong$" denotes homeomorphic.
(3) $S^m \setminus C \cong \mathbf{R}^m$.
(4) C is the monotone intersection of compact smooth m-cells.
(5) Smooth cellularity: $\mathbf{R}^m \setminus C \approx \mathbf{R}^m \setminus 0$, where "$\approx$" denotes diffeomorphic.
(6) $S^m \setminus C \approx \mathbf{R}^m$.
(7) C is properly critical and connected.

Brown showed that (1)–(3) are equivalent. Using the generalized Schoenflies theorem proved in [2], this reduces to showing that the monotone union of topological open m-cells is an open m-cell. In what follows, we extend his result to the smooth case and deduce that (4)–(7) are equivalent.

If $m \neq 4$ then there is no difference between smooth and topological open m-cells, and thus no difference between smooth and topological cellularity. This follows from the fact that $\mathbf{R}^m$ has a unique smooth structure when $m \neq 4$; see [6].

In the 4-dimensional case, however, pathology abounds: there exist compact subsets $C \subset \mathbf{R}^4$ such that $S^4 \setminus C$ is homeomorphic but not diffeomorphic to $\mathbf{R}^4$. (This amounts to the existence of "a fake $\mathbf{R}^4$ occurring as an open subset of the true $\mathbf{R}^4$," a result of A. Casson and M. Freedman; see [5, p. 98].) Such a set C is topologically cellular because it obeys (3), but is not smoothly cellular because it violates (6). Therefore, topological and smooth cellularity are inequivalent in dimension 4, and Theorem 4 becomes false if they are confounded.

SMOOTH MONOTONE CELL THEOREM. *If $X = \bigcup X_i$ is a smooth m-manifold, $X_1 \subset X_2 \subset \cdots$, and $X_i \approx \mathbf{R}^m$ for each i, then $X \approx \mathbf{R}^m$.*

Proof. Brown's construction in [3] proceeds perfectly well in the smooth case, up to the last step where one wants to assert that $X \approx \mathbf{R}^m$ follows from X being a smooth compact m-disc D to which smooth annular collars A_i are successively attached, A_1 being attached to ∂D, A_2 being attached to $\partial(D \cup A_1)$, and so on. Justifying this requires "successively straightening the smooth parameterization of A_{i+1} without disturbing that of A_i." More precisely, we establish the following lemma.

LEMMA. *Suppose that $f_1: S^{m-1} \times [0,1] \to A_1$ and $f_2: S^{m-1} \times [1,2] \to A_2$ are diffeomorphisms onto smooth m-annuli in the m-manifold X. If $A_1 \cap A_2 = f_1 S = f_2 S$, where $S = S^{m-1} \times 1$, then there is a diffeomorphism*

$$f: S^{m-1} \times [0,2] \to A_1 \cup A_2$$

extending f_1.

Proof. Set $S' = f_1 S = f_2 S$ and consider the diffeomorphism g:

$$\begin{array}{ccc} S & \xrightarrow{g} & S \\ f_1 \downarrow & & \downarrow f_2 \\ S' & \xrightarrow{=} & S'. \end{array}$$

Then $(x,s)\mapsto f_2(g(x,1),s)$ is a parameterization of A_2 that equals f_1 on S. Thus, it is no loss of generality to assume that $f_1=f_2$ on S.

Since f_1, f_2 are smooth embeddings, they extend smoothly to embeddings F_1, F_2 of $S^{m-1}\times[0,1+\delta]$ and $S^{m-1}\times[1-\delta,2]$, respectively. The product tubular neighborhood ν of S in $S^{m-1}\times\mathbf{R}$ is sent by F_1, F_2 to smooth tubular neighborhoods ν_1, ν_2 of S' in X. (See Figure 12.) By the uniqueness of tubular neighborhoods, there is a diffeomorphism $G: X\to X$ such that $G(\nu_2)=\nu_1$.

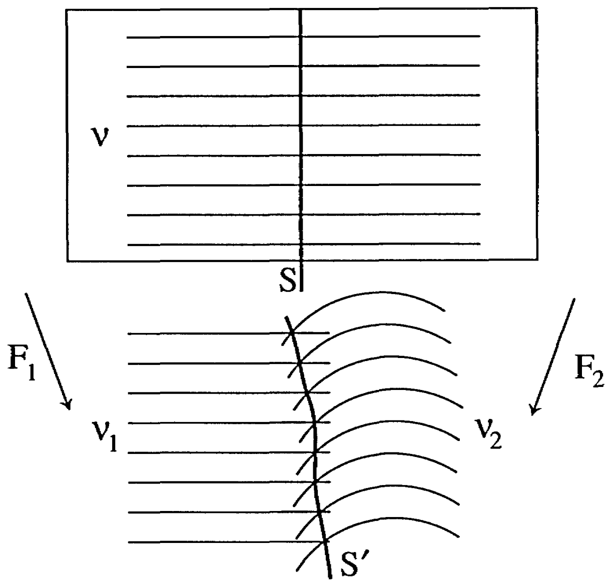

Figure 12

We may construct G so that it fixes all points on S' and all points off a small neighborhood of S'. Thus, $GA_1=A_1$, $GA_2=A_2$, and $G\circ F_2$ is a parameterization of A_2 having the same effect on ν as does F_1. That is, we may assume that $F_1(x,s)=F_2(x,\sigma)$, where $\sigma=\sigma(x,s)$ is a smooth function defined for $|1-s|\le\delta$ for which $\sigma(x,1)=1$ and $d\sigma/ds>0$. We replace σ with a smooth function $\tau=\tau(x,s)$ so that $\tau=\sigma$ on a neighborhood of $S^{m-1}\times[1-\delta,1]$, $\tau=s$ on a neighborhood of $S^{m-1}\times[1+\delta,2]$, and $d\tau/ds>0$. (See Figure 13.)

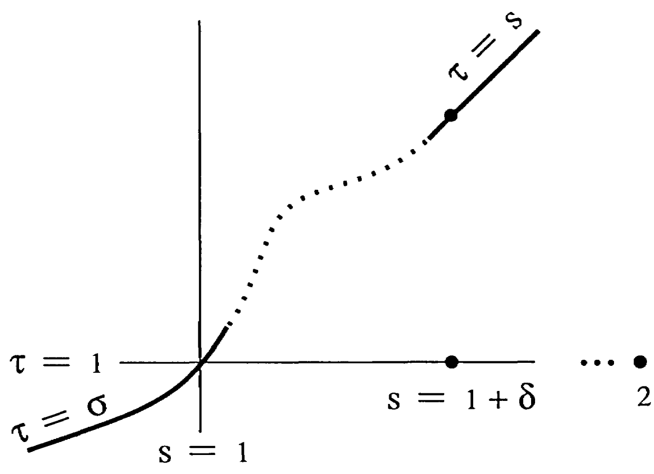

Figure 13

Then $\varphi: (x,s) \mapsto (x, \tau(x,s))$ is a diffeomorphism $S^{m-1} \times [1,2] \circlearrowleft$ and $F_1 = F_2 \circ \varphi$ near S. That is, $f_1 \cup F_2 \circ \varphi$ extends f_1 to a diffeomorphism

$$S^{m-1} \times [0,2] \to A_1 \cup A_2,$$

proving the lemma. □

Returning to the proof of the smooth monotone cell theorem, we see that a diffeomorphism $f: \mathbf{R}^m \to X = D \cup \bigcup A_i$ can be constructed as follows. Let $f_0: \mathbf{R}^m(1) \to D$ be a parameterization of D, $f_1: S^{m-1} \times [1,2] \to A_1$ a parameterization of A_1, and so forth. We may assume that $A_i \cap A_{i+1} = f_i(S_i) = f_{i+1}(S_i)$, where $S_i = S^{m-1} \times (i+1)$, $i = 1, 2, \ldots$. We identify the unit $(m-1)$-sphere, which is the boundary of $\mathbf{R}^m(1)$, with $S^{m-1} \times 1$. By the lemma, we modify f_1 to g_1 so that it extends f_0. Next we modify f_2 to g_2 so that it extends g_1, and so on. Then $f = f_0 \cup g_1 \cup g_2 \cup \cdots$ is the required diffeomorphism. □

COROLLARY. *If $m \neq 4$ then conditions* (1)–(7) *are equivalent. If $m = 4$ then the first three and the last four are equivalent, the former group being strictly weaker than the latter.*

Proof. As stated previously, equivalence of (1)–(3) was proved in [3]. In Section 1 we used the smooth monotone cell theorem to show that (5) and (7) are equivalent.

We assert that if D is a smooth compact m-cell in S^m, then its complement is a smooth open m-cell D' and the closure of D' is a compact smooth m-cell. This is clear if D is a round disc. (See Figure 14.) In general, one can use the smooth radial structure of D to define a diffeomorphism $\varphi: S^m \circlearrowleft$ fixing all points off a small neighborhood of D and shrinking D to a smooth round

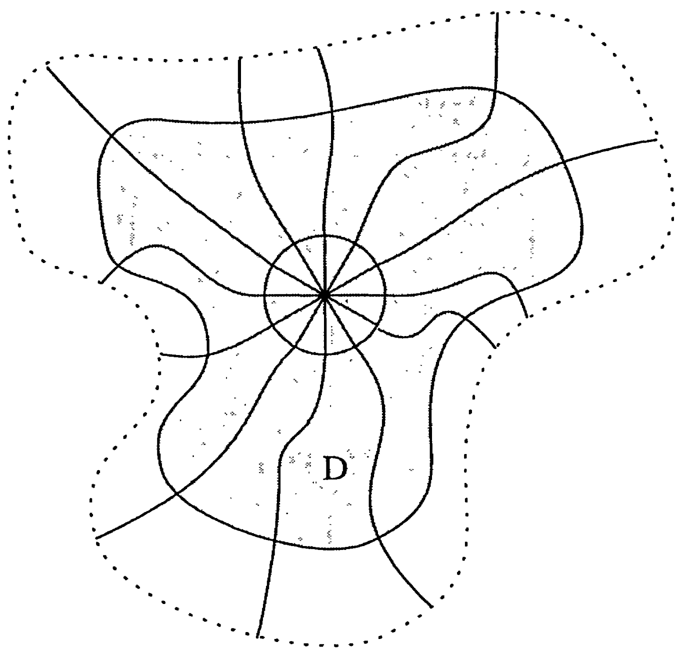

Figure 14

disc. The truth of the assertion is invariant under the ambient diffeomorphism φ. From this it follows at once that (4)$\Leftrightarrow$(6).

It is clear that (6)$\Rightarrow$(5), and so to complete the proof of the equivalence of (4)–(7) it suffices to show that (7)$\Rightarrow$(4). This is easy. Choose a large smooth disc D in $\mathbf{R}^m$ containing C in its interior, and choose a large value of T so that $\Psi(D)$ is interior to D, where $\Psi = \varphi_{-T}$ and φ is the grad(f)-flow supplied by (7). Then C is the monotone intersection of the smooth discs $\Psi^n(D)$. Thus, in all dimensions (4)–(7) are equivalent; clearly (6)$\Rightarrow$(3). If $m \neq 4$ then $\mathbf{R}^m$ has a unique smooth structure, so (3)$\Rightarrow$(6) and all the conditions are equivalent. If $m = 4$ then the Casson–Freedman fake $\mathbf{R}^4$ shows that (3) $\not\Rightarrow$ (6), and so the equivalent conditions (1)–(3) are strictly weaker than (4)–(7). □

References

1. R. Abraham and J. Robbin, *Transversal mappings and flows,* Benjamin, New York, 1967.
2. M. Brown, *A proof of the generalized Schoenflies theorem,* Bull. Amer. Math. Soc. 66 (1960), 74–76.
3. ———, *The monotone union of open n-cells is an open n-cell,* Proc. Amer. Math. Soc. 12 (1961), 812–814.
4. M. Grayson and C. Pugh, *Critical sets in* $\mathbf{R}^3$, Inst. Hautes Études Sci. Publ. Math., to appear.
5. R. Kirby, *The topology of 4-manifolds,* Lecture Notes in Math., 1374, Springer, New York, 1989.
6. R. Kirby and L. Siebenmann, *Foundational essays on topological manifolds, smoothings, and triangulations,* Ann. of Math. Stud., 88, Princeton Univ. Press, Princeton, N.J., 1977.

Alec Norton
Department of Mathematics
University of Texas at Austin
Austin, TX 78712

Charles Pugh
Department of Mathematics
University of California, Berkeley
Berkeley, CA 94720