PROCEEDINGS OF THE
AMERICAN MATHEMATICAL SOCIETY
Volume 98, Number 2, October 1986

EVERY EXACTLY 2-TO-1 FUNCTION ON THE REALS HAS AN INFINITE SET OF DISCONTINUITIES

JO HEATH

ABSTRACT. It has long been known that the set of discontinuities of a 2-to-1 function on either the closed or the open interval must be nonempty; this paper proves that the set must be infinite.

Beginning in 1939 with O. G. Harrold's proof [4] that every (exactly) 2-to-1 function on $[0,1]$ must be discontinuous, numerous mathematicians have considered 2-to-1 functions (see [1–9]). In particular, it was shown in [7, 9, 2, and 3], respectively, that only discontinuous 2-to-1 functions exist on the reals, on the closed 2-cell, on the closed 3-cell, or on any 1-complex with odd Euler characteristic. None of the papers [1] through [9] consider whether the set of discontinuities must be infinite, and recently R. Levy raised the question answered by the title.

This paper shows that the set of discontinuities must be infinite for any 2-to-1 function defined on $[0,1]$ (Theorem 1), on the reals (Corollary 1), or on any 1-complex with odd Euler characteristic (Theorem 2). Krystyna Kuperberg has recently constructed an exactly 2-to-1 function on a 2-cell with a single discontinuity. This map easily extends to a map on an n-cell ($n > 2$) with the same properties.

THEOREM 1. *There is no function from* $[0,1]$ *to a Hausdorff space that is exactly* 2*-to-*1 *and has only a finite number of discontinuities.*

PROOF. Suppose f is such a function. Define g: $[0,1] \to [0,1]$ by $g(x) = y$ iff $x \neq y$ and $f(x) = f(y)$. Then g satisfies the following properties:

(i) $g = g^{-1}$, g is 1-to-1, g has no fixed point, and

(ii) if $\{x_i\} \to x$, $\{g(x_i)\} \to w$ and both x and w are points of continuity of f, then either $w = x$ or $w = g(x)$. Property (ii) follows from the fact that the image of f is Hausdorff and the fact that $w = x$ or $w = g(x)$ is equivalent to $f(x) = f(w)$.

It will be shown that no such g exists.

Denote by DIS the finite set of discontinuities of f. Then if x is not in DIS and x is a point of discontinuity of g, x will belong to either $S1$ or $S2$ or both:

DEFINITION. The point x in $[0,1]$ belongs to S1 iff there is a sequence $\{x_i\}$ converging to x such that $\{g(x_i)\}$ converges to a point of DIS.

Received by the editors July 15, 1985 and, in revised form, September 18, 1985 and October 9, 1985.
Presented 4/4/86 Spring Topology Conference, University of Southwestern Louisiana.
1980 *Mathematics Subject Classification* (1985 *Revision*). Primary 54C10; Secondary 26A03, 26A15.

0002-9939/86 $1.00 + $.25 per page

DEFINITION. The point x in $[0,1]$ belongs to $S2$ iff there is a sequence $\{x_i\}$ converging to x such that $\{g(x_i)\}$ converges to x.

The sets $S1$ and $S2$ are closed. Suppose $\{y_i\} \to y$ and each y_i is in S1. For each i there is a sequence of points converging to y_i whose g images converge to some w_i in DIS. Choose a number x_i from this sequence within $1/i$ of y_i such that $g(x_i)$ is within $1/i$ of w_i. Since DIS is finite, it can be assumed that all of the w_i are equal to the same w in DIS. Thus the $\{x_i\}$ sequence converges to y and the g images converge to w, so y is in S1. In a similar way, $S2$ is closed.

Let $S = \text{DIS} \cup S1 \cup S2$, and let C denote the complementary (open) intervals of $[0,1] - S$. Note that there may be a finite number of points in S at which g is coincidentally continuous, but if g is discontinuous at x then x is in S.

Some of the discontinuities of g are endpoint discontinuities and the following classify the two types:

DEFINITION. The point x belongs to $E1$ iff for some interval I in C, x is an endpoint of I and there is a sequence of points in I that converges to x and whose g images converge to a point of DIS. Note that $E1 \subseteq S1$.

DEFINITION. The point x belongs to $E2$ iff for some interval I in C, x is an endpoint of I and there is a sequence of points in I that converges to x and whose g images converge to x. Note that $E2 \subseteq S2$.

Consider the graph of g near a point b in $E2$ that is the endpoint of the interval (a, b) of C. The interval (a, b) misses S, so g is continuous and 1-to-1 on (a, b) and the graph of g is either an increasing or a decreasing open arc A over (a, b). From the definition of $E2$, the right endpoint of A is the point (b, b), which is not on the graph of g. Since $g = g^{-1}$, g is symmetric with respect to the diagonal line $\{x = y\}$, and so there is a symmetric arc A' in the graph of g over (b, c). The arc A must be decreasing (and A' too) because if A were increasing over (a, b) up to the point (b, b) then some vertical line would intersect both A and A'. Thus neither the left endpoint of A nor the right endpoint of A' can lie on the diagonal $\{x = y\}$ implying that neither a nor c is in $E2$. Hence,

FACT 1. If b is in $E2$ then there is an interval (a, c) containing b such that

(1) (a, c) contains no other point of discontinuity of g,

(2) neither a nor c belongs to $E2$, and

(3) if a and c also do not belong to $E1$ then the points of the graph of g at a and c, $(a, g(a))$ and $(c, g(c))$, are symnmetric images about the diagonal line $\{x = y\}$.

Property (3) is true because g is continuous from the right at a and continuous from the left at c, and the symmetric points would be endpoints of A and A'.

The set $E1$ is finite. Suppose b is an endpoint of an interval (a, b) in C satisfying the definition of $E1$. Again the graph of g over (a, b) is an open arc A. The right endpoint of A is the point (b, w) for some point w in DIS. The interval (a, b) maps then either onto some interval (c, w) or onto some interval (w, c) depending on whether g is increasing or decreasing on (a, b). Since g is 1-to-1, only one interval can map onto any interval with right endpoint w and only one can map onto any interval with left endpoint w. Thus $E1$ is finite because DIS is.

Now let $E1'$ and DIS$'$ denote the points of $E1$ and $\text{DIS} \cup \{x \mid g(x) \in \text{DIS}\}$, respectively, that are not limit points of S.

Define $S' = S - (\mathrm{DIS}' \cup E1' \cup E2)$. No point of $E2$ is a limit point of S (see Fact 1, part (1)), and S is closed, so S' is closed. It will now be shown that S' is in fact empty. Note that if S' is empty, then S is equal to a set $E2$ of isolated (in S) points plus a finite set $E1' \cup \mathrm{DIS}'$ that contains no limit point of $E2$; and then, since S is closed, $E2$ and hence S is finite.

Let $Q1, Q2, \ldots$ denote a sequence of closed and convex quadralaterals that miss the diagonal $\{x = y\}$ and each of the finitely many lines $\{x = w\}$ and $\{y = w\}$, for all $w \in \mathrm{DIS}$, such that the sequence covers the rest of the square $[0,1] \times [0,1]$. Since the graph of g misses the diagonal (no fixed points), the union of the quadralaterals contains all of the graph of g except for a finite set. For each natural number i define Ki to be the set of points in S' whose graph point lies in Qi. For completeness, define $K0$ to be the points of S' that are either in DIS themselves or whose g image is in DIS. Then S' is the union of the Kj, for j nonnegative.

First, each Ki is closed. $K0$ is finite, and if $i > 0$ suppose that $\{x_j\}$ is a sequence of points in Ki. Each graph point $(x_j, g(x_j))$ is in Qi and some subsequence of these graph points converges to the point (x, y) in Qi. Since S' is closed, x is in S'. If $g(x) \neq y$, then either x is in DIS, y is in DIS, or $x = y$ (see property (ii) of g). But Qi was designed to miss such points. Hence $g(x) = y$, x is in Ki, and Ki is closed. For further reference note that the following fact has also been shown:

FACT 2. The restriction of g to Ki is continuous, for each i.

Second, each Ki is nowhere dense (in S'). On the contrary, suppose some Ki, $i > 0$, fails to be nowhere dense. There is a open interval I in $[0,1]$ such that $I \cap S' \subset Ki$, and $I \cap S'$ is nonempty. Since $E1$ and DIS are finite and their isolated (in S) points removed from S, there is a point x in $I \cap S'$ that is not in $E1 \cup \mathrm{DIS}$, and a smaller interval I' containing x and no point of $E1 \cup \mathrm{DIS}$. From the definition of S', x is not in $E2$ either, and so x is not an endpoint discontinuity of g. This means x is a limit point of $S1 \cup S2$. There cannot be a subinterval of I contained in $S1 \cup S2$ because g restricted to Ki is continuous (see Fact 2). Hence there are at least three intervals $I1$, $I2$, $I3$ of C that lie in I'. Consider $I2 = (a, b)$. Assume b is in $E2$. From Fact 1 there is an interval (a, c) containing b and no other discontinuity of g (so (a, c) must lie in I') and such that the graph points $(a, g(a))$ and $(c, g(c))$ are symmetric images about the diagonal $\{x = y\}$. But a and c in Ki means both graph points are in Qi, and, because Qi is convex, Qi intersects the diagonal. This contradiction implies that b is not in $E2$. Since b is not in $E1$ either, g is continuous on $(a, b]$. The argument for the left endpoint is the same, so g is continuous on the closure of every interval of C that lies in I' with the possible exception of the first and last. Since g restricted to Ki is also continuous (Fact 2) and g is one-to-one, g is continuous on some interval containing x. This contradicts the fact that x is a limit point of $S1 \cup S2$ and $S1 \cup S2$ contains only finitely many points of continuity of g.

This means that S' is a complete (S' is closed in $[0,1]$) metric space and is the countable union of closed, nowhere dense (in S') subsets.

Hence S' is empty and S is finite.

The graph of g then consists of a finite set A of closed, open, or half-closed and half-open arcs plus a finite set P of isolated points. First, P must be even since if

(x, y) is in P then (y, x) is also in P (because $g = g^{-1}$) and since g contains no point on the diagonal. Second, the number of closed intervals in A and the number of open intervals in A must also be even because of the symmetry of g. Now suppose that M is a set of disjoint closed, open, and half-closed and half-open intervals of least cardinality such that

1. $[0, 1]$ is the union of the elements of M plus an even number of points disjoint from the elements of M,

2. the number of closed intervals in M is even, and

3. the number of open intervals in M is even.

The final contradiction is that no such M exists.

First, M contains no closed interval. If $[a, b]$ is in M there is another $[c, d]$ in M; assume that $b < c$. The interval next to $[a, b]$ to the right in M has the form $(b, x\}$, where "}" is a wild card symbol for either ")" or "]", and the interval of M just to the left of $[c, d]$ has the form $\{y, c)$; there is also the possibility that (b, c) belongs to M. If $[a, b]$ and $[c, d]$ are deleted from M and $(b, x\}$ and $\{y, c)$ replaced by $(a, x\}$ and $\{y, d)$ respectively, or in the latter case (b, c) is replaced by (a, d), the smaller interval collection has the same properties, contradicting the minimality of M.

Second, there is no interval $[a, b)$ in M with $a \neq 0$ (nor is there an $(a, b]$ with $b \neq 1$). If there were, the interval just to the left would have the form $\{c, a)$. If $[a, b)$ and $\{c, a)$ are deleted from M and $\{c, b)$ added to M a smaller set with the same properties is formed.

Third, there do not exist two adjacent open intervals, (a, b) and (b, c), in M. There is a third interval in M since $(0, b)$ and $(b, 1)$ leave three points uncovered. Suppose then that $(c, d\}$ is in M. Again replace (a, b), (b, c) and $(c, d\}$ with $(a, d\}$ for a smaller set with the same properties.

This leaves only the impossibility that M contains only $[0, b)$ and $(b, 1]$.

This proves Theorem 1.

LEMMA 1. *Suppose $A = B \cup F$ where F misses B and contains an even (finite) number of points and each of A and B is a T_1-space. Then there is an exactly 2-to-1 function on A with a Hausdorff image and with at most finitely many discontinuities iff there is one on B.*

COROLLARY 1. *There is no exactly 2-to-1 function on the reals with Hausdorff image that has only finitely many discontinuities.*

PROOF. The reals $\cong (0, 1) = [0, 1] - \{0, 1\}$.

THEOREM 2. *There is no exactly 2-to-1 function on a 1-complex with odd Euler characteristic with Hausdorff image and with only finitely many discontinuities.*

PROOF. Suppose f is such a function on the 1-complex X. Then X is the union of i open disjoint intervals (the edges minus their endpoints), $e1, e2, \ldots, ei$, and a disjoint set with j points (the vertices), $v1, v2, \ldots, vj$, where $j - i$ is odd.

For each $k = 1, 2, \ldots, i$ let g_k denote a homeomorphism from the interval $(k, k + 1)$ on the real line onto ek.

CASE 1. $j > i$.

Define f' on $[1, i+1] \cup \{i+2, i+3, \ldots, j\}$ by $f'(x) = f(g_k(x))$ if x is in $(k, k+1)$ and $f'(k) = f(vk)$. The new function f' may introduce new discontinuities at the integers (vertices), but f' is still a 2-to-1 function on $[1, i+1]$ plus an even (possibly zero) number of points with only finitely many discontinuities. This contradicts Lemma 1 and Theorem 1.

CASE 2. $i > j$.

Define f' on $[1, i+1] - \{j+1, j+2, \ldots, i+1\}$ by $f'(x) = f(g_k(x))$ if x is in $(k, k+1)$ and $f'(k) = f(vk)$. Then f' is the same type of function on $[i, i+1]$ minus an even number of points, again contradicting Lemma 1 and Theorem 1.

REFERENCES

1. K. Borsuk and R. Molski, *On a class of continuous maps*, Fund. Math. **45** (1958), 84–98.

2. P. Civin, *Two-to-one mappings of manifolds*, Duke Math. J. **10** (1943), 49–57.

3. P. Gilbert, *n-to-one mappings of linear graphs*, Duke Math. J. **9** (1942), 475–486.

4. O. G. Harrold, *The non-existence of a certain type of continuous transformation*, Duke Math. J. **5** (1939), 789–793.

5. _____, *Exactly* $(k,1)$ *transformations on connected linear graphs*, Amer. J. Math. **62** (1940), 823–834.

6. V. Martin and J. H. Roberts, *Two-to-one transformations on 2-manifolds*, Trans. Amer. Math. Soc. **49** (1941), 1–17.

7. J. Mioduszewski, *On two-to-one continuous functions*, Dissertationes Math. (Rozprawy Mat.) **24** (1961), 42.

8. S. B. Nadler, Jr., and L. W. Ward, Jr., *Concerning exactly* $(n,1)$ *images of continua*, Proc. Amer. Math. Soc. **87** (1983), 351–354.

9. J. H. Roberts, *Two-to-one transformations*, Duke Math. J. **6** (1940), 256–262.

DEPARTMENT OF MATHEMATICS, AUBURN UNIVERSITY, AUBURN, ALABAMA 36849

Current address: Department of Mathematics, The University of Reading, Whiteknights, PO Box 220, Reading RG6 2AX, England