LECTURE 7: SARD'S THEOREM

1. Critical points and critical values

¶ Critical points and critical values: the definition.

Recall that in calculus,

- a number $a \in \mathbb{R}$ is called a critical point of a smooth *function* $f : \mathbb{R} \to \mathbb{R}$ if $f'(a) = 0$.
- a point $a \in \mathbb{R}^m$ is called a critical point of a smooth *function* $f : \mathbb{R}^m \to \mathbb{R}$ if $\frac{\partial f}{\partial x^i} = 0$ for all i (i.e., the differential $df_a : \mathbb{R}^m \to \mathbb{R}$ is not surjective).
- a point $a \in \mathbb{R}^m$ is a critical point of a smooth *map* $f : \mathbb{R}^m \to \mathbb{R}^n$ between Euclidean open sets if $df_a : \mathbb{R}^m \to \mathbb{R}^n$ is not surjective (i.e., if f is not a submersion at a).

This conception can be easily extended to smooth maps between smooth manifolds:

Definition 1.1. Let M, N be smooth manifolds and $f : M \to N$ a smooth map.

(1) We say $p \in M$ is a *critical point* of f if $df_p : T_pM \to T_{f(p)}N$ is not surjective[1]. We say $p \in M$ is a *regular point* of f if it is not a critical point.
(2) We say $q \in N$ is a *regular value* of f if any $p \in f^{-1}(q)$ is a regular point. We say $q \in N$ is a *critical value* of f if it is not a regular value.

We will denote the set of all critical points of f by $\mathrm{Crit}(f)$.

Remark. Let $f : M \to N$ be a smooth map.

- By definition, any $q \in N \setminus \mathrm{Im}(f)$ is automatically a regular value. [So a regular value of a smooth map need not be a value of the map!]
- Critical values are exactly the image of critical points, but the pre-image of critical values may contain regular points. [So the image of a regular point could be a critical value!]
- By definition, the set of regular points is open in M:
  $$\begin{aligned} &p \text{ is a regular point of } f \\ \iff &f \text{ is a submersion at } p \\ \iff &f \text{ is a submersion near } p, \text{ i.e. points “near” } p \text{ are regular points.} \end{aligned}$$
  As a consequence, the set of critical points is closed in M. However, the set of critical values need not be closed in N, and the set of regular values need not be open in N.

Critical/regular values are important in the study of smooth maps/manifolds. For example, we will show that if q is a regular value of a smooth map $f : M \to N$, then $f^{-1}(q)$ is a *smooth submanifold* (the conception will be defined later) of M.

[1] We define critical points to be those points such that $\mathrm{rank}(df_p) < \dim N$. In some books critical points are defined to be those points such that $\mathrm{rank}(df_p) < \min(\dim M, \dim N)$.

¶ Examples.

In calculus we have seen that any max/min value points are critical. This is still true for smooth functions defined on smooth manifolds:

Proposition 1.2. *Let $f \in C^\infty(M)$ be a smooth function and $p \in M$ be a maximal or minimal value point of f. Then p is a critical point of f. [As a consequence, any smooth function defined on a compact manifold admits at least two critical points.]*

Proof. For any $X_p = \sum a_i \partial_i|_p \in T_pM$, we have

$$X_p(f) = \sum a_i \partial_i|_p(f) = \sum a_i \frac{\partial(f \circ \varphi^{-1})}{\partial x^i}(\varphi(p)) = 0.$$

So when applied to the function $g(t) = t$ (viewed as an element in $C^\infty(\mathbb{R})$), we get

$$df_p(X_p)(g) = X_p(g \circ f) = X_p(f) = 0.$$

This implies $df_p(X_p) = 0$ (since if we write $df_p(X_p) = c\frac{d}{dt}$, then $c = c\frac{d}{dt}g$). So p is a critical point of f. □

Note that in the argument above, we really proved the following useful lemma:

Lemma 1.3. *If we identify $T_t\mathbb{R}$ with $\mathbb{R}$ [by identifying $c\frac{d}{dt}$ with c], then*

$$df_p(X_p) = X_p(f), \qquad \forall f \in C^\infty(M), \forall X_p \in T_pM,$$

Example. As a consequence of Proposition 1.2, the "height function"

$$f : S^n \to \mathbb{R}, \qquad f(x^1, \cdots, x^{n+1}) = x^{n+1}.$$

has the north pole $(0, \cdots, 0, 1)$ and the south pole $(0, \cdots, 0, -1)$ as its critical points f. It is not hard to check that all other points in S^n are regular points of f (check!). Thus the only critical values of f are 1 and -1.

Example. For the following two extremal cases,

- $f : M \to N$ is a constant map, i.e. $f(p) \equiv q_0 \in N$,
- $f : M \to N$ is any smooth map, but $\dim M < \dim N$,

any point in M is a critical point, and thus any point in $f(M)$ is a critical value of f.

Example. Here is a function $f \in C^\infty(\mathbb{R})$ whose regular values are dense in $\mathbb{R}$: we can list all rational numbers as $\mathbb{Q} = \{r_1, r_2, \cdots\}$. Then we take a smooth bump function f_0 defined on $\mathbb{R}$ such that

$$\text{supp}(f_0) \subset (-1/3, 1/3) \quad \text{and} \quad f_0 \equiv 1 \text{ on } (-1/4, 1/4).$$

Let

$$f(x) = \sum_{k=1}^\infty r_k f_0(x - k).$$

Then each $k \in \mathbb{N}$ is a critical point of f, and thus the set of critical values of f contains $f(\mathbb{N}) = \mathbb{Q}$. (Question: Can you find an example whose critical values form an uncountable set?)

2. Sard's Theorem

¶ Measure zero sets on smooth manifolds.

The main theorem we will prove today is: For any smooth map $f : M \to N$, "most" points in N (the complement of a negligible set in the sense of measure) are regular values of f.

Let's first explain the phrase "negligible in the sense of measure", namely "measure zero sets" on smooth manifolds. Note that we have not introduce any measure on M or N yet. However, it turns out that we don't need to define a measure when we are only talking about "measure zero sets" on smooth manifolds! The idea is: use the Lebesgue measure on Euclidean space.

We should be careful when we are trying to define something on manifolds via the corresponding Euclidean ones: We may want to "transplant" the Lebesgue measure on the Euclidean space to manifolds by using local charts. However, this does not give us a well-defined measure on manifolds since it depends on the choice of local charts. In fact, a measure structure is an extra structure on manifold. With only a smooth structure at hand, we don't have a canonical choice of measure structure. (We will see that for an orientable manifold, each volume form gives rise to a measure).

However, "whether a set is of measure zero or not" make sense without introducing a measure: again we use the Lebesgue measure on Euclidian space, but now it is independent of the choices of local charts. Recall that a subset $A \subset \mathbb{R}^n$ is of *measure zero* if for any $\varepsilon > 0$, there exists a countable union of open boxes $U_i \in \mathbb{R}^n$ so that

$$A \subset \bigcup_i U_i \qquad \text{and} \qquad \sum_i \text{volume}(U_i) < \varepsilon.$$

For Lebesgue measure zero sets in $\mathbb{R}^n$, we have the following properties:

(i) A countable union of measure zero sets is a measure zero set.
(ii) If $A \subset \mathbb{R}^n$ is a measure zero set, and $f : \mathbb{R}^n \to \mathbb{R}^n$ is smooth, then $f(A)$ is a measure zero set in $\mathbb{R}^n$. [2]

Since any manifold M can by covered by countable many charts, and each chart identify an open set in M with an open set in $\mathbb{R}^n$, the following definition is independent of coordinate charts (and thus is reasonable):

Definition 2.1. We say $A \subset M$ is a *measure zero* set if for any $p \in A$, one can find a chart (φ, U, V) of M near p so that $\varphi(A \cap U)$ is a measure zero set in V.

It is also obvious that a countable union of measure zero sets in a smooth manifold is still a measure zero set. We will need the following special case of Fubini's theorem:

Theorem 2.2 (Fubini's Theorem — a special case)**.** *Let A be a measurable subset of $\mathbb{R}^n$ such that the "slice" $A \cap (\{c\} \times \mathbb{R}^{n-r})$ has Lebesgue measure zero in $\mathbb{R}^{n-r}$ for all $c \in \mathbb{R}^r$. Then A has Lebesgue measure zero in $\mathbb{R}^n$.*

[2] In real analysis we know that a continuous function could map a measure zero set in $\mathbb{R}^n$ to a set with positive measure in $\mathbb{R}^n$. However, a (local) Lipschitz map will always map a measure zero set in $\mathbb{R}^n$ to a measure zero set in $\mathbb{R}^n$.

¶ Sard's theorem.

Now we are ready to state and prove the following remarkable theorem in differential topology, which claim that the set of critical values is negligible in N. [However, the theorem does not claim that the set of critical points is a measure zero subset in M. In fact as we have seen, it could have that all points in M are critical points.]

Theorem 2.3 (Sard's Theorem[3]). *For any smooth map $f : M \to N$, the set of all critical values of f is of measure zero in N.*

Note that if $n = \dim N = 0$, then there is no critical points (and thus no critical values) at all. So in what follows we may assume $n > 0$.

Since we may cover M by at most countable many coordinate charts U_i, with each $f(U_i) \subset X_i$ for some coordinate charts X_i of N, and since a countable union of measure zero sets is still of measure zero, it suffices to prove the theorem for smooth maps between Euclidean open sets, i.e.

Theorem 2.4. *If $U \subset \mathbb{R}^m$ and $X \subset \mathbb{R}^n$ are Euclidean open subsets, and $f : U \to V$ is smooth, then the set of all critical values is of measure zero in $\mathbb{R}^n$..*

Proof. First observe that if $m < n$, then the theorem holds trivially. In fact, in this case we can prove that the whole image $f(U)$ is of measure zero in $\mathbb{R}^n$. To see this, we identify U with the subset $U \times \{0\}$ inside $U \times \mathbb{R}^{n-m} \subset \mathbb{R}^n$. Obviously $U \times \{0\}$ has measure zero in $\mathbb{R}^n$. Now define a map $\tilde{f} : U \times \mathbb{R}^{n-m} \to \mathbb{R}^n$ via $\tilde{f}(x, y) = f(x)$. Then $f(U) = \tilde{f}(U \times \{0\})$. ($\tilde{f}$ is smooth since it is the composition of two smooth maps: the projection map $\pi : U \times \mathbb{R}^{n-m} \to U$ and $f : U \to \mathbb{R}^n$.) Since $\tilde{f}$ is a smooth between Euclidean spaces of the same dimension, and $U \times \{0\}$ is of measure zero in $\mathbb{R}^n$, we conclude that the image $\tilde{f}(U \times \{0\})$ is of measure zero in $\mathbb{R}^n$.

In what follows we will proceed by induction. The theorem is certainly true for $m = 0$, since any countable set has measure zero. We will proceed to prove that the theorem is true for m assuming that it is true for $m-1$. Let C be the set of all critical points of f, then we need to show that $f(C)$ is of measure zero in N. Denote

$$C_j = \{x \in U \mid \partial^\alpha f(x) = 0 \text{ for all } |\alpha| \le j\}.$$

Obviously for any positive integer k,

$$f(C) = f(C \backslash C_1) \cup f(C_1 \backslash C_2) \cup \cdots \cup f(C_{k-1} \backslash C_k) \cup f(C_k).$$

Following J. Milnor, we will divide the proof into three steps:

Step 1: $f(C \backslash C_1)$ has measure zero.
Step 2: $f(C_i \backslash C_{i+1})$ has measure zero for each i.
Step 3: $f(C_k)$ has measure zero for large k, say for $k \ge \frac{m}{n}$.

[Question: Why do we need Step 3? Are Steps 1 and 2 enough?]

[3] The theorem was first proven by Morse in 1939 for $m = 1$, and was extended to the general case by Sard in 1942. A version for infinite-dimensional Banach manifolds was proven by Stephen Smale in 1965.

Proof of step 1. For each $x \in C\setminus C_1$, we will find an open set $U_x \ni x$ such that $f(U_x \cap C)$ has measure zero. Since $C\setminus C_1$ can be covered by countably many such open sets (by second-countability), this implies $f(C\setminus C_1)$ is of measure zero.

Note that if $n = 1$, then a point x is a critical point of f if and only if $\frac{\partial f}{\partial x^i}(x) = 0$ for all i. It follows $C\setminus C_1 = \emptyset$, so $f(C\setminus C_1)$ must be of measure zero. So we may assume $m \geq n > 1$. Since $x \notin C_1$, there is some partial derivative, say $\frac{\partial f_1}{\partial x^1}$, is not zero at x. Consider

$$h : U \to \mathbb{R}^m, \quad h(x) = (f_1(x), x^2, \cdots, x^m).$$

Then dh_x is non-singular. According to the inverse function theorem, h maps a neighborhood U_x of x diffeomorphically onto an open set V in $\mathbb{R}^m$. The composition $g = f \circ h^{-1}$ will then map V into $\mathbb{R}^n$. Moreover, since h^{-1} is a diffeomorphism on V, dh^{-1} is a linear isomorphism everywhere in V. So the set of critical values of g is exactly $f(U_x \cap C)$.

Note that the map g we constructed is of the form (why?)

$$g(t, x^2, \cdots, x^m) = (t, g_2, \cdots, g_n).$$

So for each t, g induces a smooth map $g^t : (\{t\} \times \mathbb{R}^{m-1}) \cap V \to \{t\} \times \mathbb{R}^{n-1}$. Moreover,

$$dg = \begin{pmatrix} 1 & 0 \\ * & \left(\frac{\partial (g^t)_i}{\partial x^j}\right)_{i,j\geq 2} \end{pmatrix}.$$

It follows that a point in $(\{t\} \times \mathbb{R}^{m-1}) \cap V$ is critical for g^t if and only if it is critical for g. But by the induction hypothesis, Sard's theorem is true for $m - 1$, i.e. holds for each g^t. So the set of critical values of g^t has measure zero in $\{t\} \times \mathbb{R}^{n-1}$. Finally by applying Fubini's theorem, we see that the set of critical values of g is of measure zero.

Proof of step 2. For each $x \in C_i\setminus C_{i+1}$, one can find some multi-index α with $|\alpha| = i$, so that

- the partial derivative $w := \partial^\alpha f$ vanishes on C_i,
- at least one first order partial derivative of w, say $\frac{\partial w}{\partial x^1}$, does not vanish at x.

Again by applying the inverse function theorem, we conclude that

$$h : U \to \mathbb{R}^m, \quad h(x) = \big(w(x), x^2, \cdots, x^m\big)$$

maps a neighborhood U_x of x diffeomorphically onto an open set V in $\mathbb{R}^m$. (To get a better understand and/or to avoid possible mistakes, you may want to think about the meaning of this map for $m = 1$.) By construction, h carries $C_i \cap U_x$ into the hyperplane $\{0\} \times \mathbb{R}^{m-1}$. Again we consider the map $g = f \circ h^{-1}$. Then the critical points of g of type C_i are all in the hyperplane $\{0\} \times \mathbb{R}^{m-1}$. Let

$$\bar{g} : (\{0\} \times \mathbb{R}^{m-1}) \cap V \to \mathbb{R}^n$$

be the restriction of g. Then the set of critical points of g of type C_i coincides with the set of critical points of $\bar{g}$. By induction, the set of critical values of $\bar{g}$ is of measure zero in $\mathbb{R}^n$. It follows that the image of the critical points of g of type C_i is of measure zero. Therefore, $f(C_i \cap U_x)$ is of measure zero. Since $C_i\setminus C_{i+1}$ can be covered by countable many such sets U_x, $f(C_i\setminus C_{i+1})$ is of measure zero.

Proof of step 3. Let $Q \subset U$ be a cube whose sides are of length δ. We will prove that for $k > \frac{m}{n} - 1$, $f(C_k \cap Q)$ has measure zero. Since C_k can be covered by countably many such cubes, this implies that $f(C_k)$ has measure zero.

From Taylor's theorem, the compactness of Q and the definition of C_k, we see that

$$f(x+h) = f(x) + R(x,h),$$

where $|R(x,h)| < a|h|^{k+1}$ for $x \in C_k \cap Q, x+h \in Q$, and the constant a depends only on f and Q. Now we subdivide Q into r^m cubes whose sides are of length $\frac{\delta}{r}$. Let Q_1 be a cube of subdivision that contains a point $x \in C_k$. Then any point of Q_1 can be written as $x+h$ with $|h| < \sqrt{m}\frac{\delta}{r}$. It follows that $f(Q_1)$ lies in a cube with sides of length $\frac{b}{r^{k+1}}$ centered about $f(x)$, where $b = 2a(\sqrt{m}\delta)^{k+1}$ is a constant. So $f(C_k \cap Q)$ is contained in the union of at most r^m cubes having total volume

$$\mathrm{Vol} \le r^m \left(\frac{b}{r^{k+1}}\right)^n = b^n r^{m-(k+1)n}.$$

Since $k > \frac{m}{n} - 1$, we see Vol $\to 0$ as $r \to \infty$. It follows that $f(C_k \cap Q)$ is of measure zero. □

Remark. Sard's theorem holds for C^r maps between C^r manifolds, provided $r \ge 1 + \max(m-n, 0)$. Moreover, this is sharp. However, the proof above does not work in this more general setting. [Can you see why?]