ALEXANDROV'S THEOREM ON THE SECOND DERIVATIVES OF CONVEX FUNCTIONS *VIA* RADEMACHER'S THEOREM ON THE FIRST DERIVATIVES OF LIPSCHITZ FUNCTIONS

RALPH HOWARD
DEPARTMENT OF MATHEMATICS
UNIVERSITY OF SOUTH CAROLINA
COLUMBIA, S.C. 29208, USA
HOWARD@MATH.SC.EDU

These are notes from lectures given the the functional analysis seminar at the University of South Carolina.

Contents

1. Introduction

A basic result in the regularity theory of convex sets and functions is the theorem of Alexandrov that a convex function has second derivatives almost everywhere. The notes here are a proof of this following the ideas in the appendix of the article [4] of Crandall, Ishii, and Lions and they attribute the main idea of the proof to F. Mignot [5]. To make the notes more self contained I have included a proof of Rademacher's theorem on the differentiable almost everywhere of Lipschitz functions following the presentation in the book [8] of Ziemer (which I warmly recommend to anyone wanting

Date: May 1998.

to learn about the pointwise behavior of functions in Sobolev spaces or of bounded variation). Actually a slight generalization of Alexandrov's theorem is given in Theorem 5.3 which shows that set-valued functions that are inverses to Lipschitz functions are differentiable almost everywhere.

To simplify notation I have assumed that functions have domains all of $\mathbf{R}^n$. It is straightforward to adapt these proofs to locally Lipschitz functions or convex function defined on convex open subsets of $\mathbf{R}^n$.

As to notation. If $x, y \in \mathbf{R}^n$ then the inner product is denoted as usual by either $x \cdot y$ or $\langle x, y\rangle$. Explicitly if $x = (x_1, \dots, x_n)$ and $y = (y_1, \dots, y_n)$

$$x \cdot y = \langle x, y\rangle = x_1 y_1 + \cdots + x_n y_n.$$

The norm of x is $\|x\| = \sqrt{x \cdot x}$. Lebesgue measure in $\mathbf{R}^n$ will be denoted by $\mathcal{L}^n$ and integrals with respect to this measure will be written as $\int_{\mathbf{R}^n} f(x)\,dx$. If $x_0 \in \mathbf{R}^n$ and $r > 0$ then the open and closed balls about x_0 will be denoted by

$$B(x_0, r) := \{x \in \mathbf{R}^n : \|x - x_0\| < r\} \quad \overline{B}(x_0, r)\{x \in \mathbf{R}^n : \|x - x_0\| \le r\}.$$

2. Rademacher's Theorem

We first review a little about Lipschitz functions in one variable. The following is a special case of a theorem of Lebesgue.

2.1. Theorem. *Let $f\colon \mathbf{R} \to \mathbf{R}$ satisfy $|f(x_1) - f(x_0)| \le M|x_1 - x_0|$. Then the derivative $f'(t)$ exists for almost all t and*

$$|f'(t)| \le M$$

holds at all points where it does exist. Also for $a < b$

$$\int_a^b f'(t)\,dt = f(b) - f(a). \qquad \square$$

Note if $f\colon \mathbf{R} \to \mathbf{R}$ is Lipschitz and $\varphi \in C_0^\infty(\mathbf{R})$ then the product $\varphi(t)f(t)$ is also Lipschitz and so the last result implies

$$\int_{\mathbf{R}} f'(t)\varphi(t)\,dt = -\int_{\mathbf{R}} f(t)\varphi'(t)\,dt.$$

Now if $f\colon \mathbf{R}^n \to \mathbf{R}$ then denote by $D_i f$ the partial derivative

$$D_i f(x) = \frac{\partial f}{\partial x_i}(x)$$

at points where this partial derivative exists. Let $Df(x)$ denote

$$Df(x) = (D_1 f(x), \dots, D_n f(x)).$$

2.2. Proposition. *If $f\colon \mathbf{R}^n \to \mathbf{R}$ is Lipschitz, say $\|f(x_1) - f(x_0)\| \le M\|x_1 - x_0\|$, then $Df(x)$ exists for almost all $x \in \mathbf{R}^n$. Moreover all the partial derivative $D_i f$ satisfy*

$$|D_i f(x)| \le M$$

at points where they exist. Thus

$$\|Df(x)\| \le \sqrt{n}M. \tag{2.1}$$

Finally if $\varphi \in C_0^\infty(\mathbf{R}^n)$ then

$$\int_{\mathbf{R}^n} D_i f(x)\varphi(x)\,dx = -\int_{\mathbf{R}^n} f(x) D_i\varphi(x)\,dx. \tag{2.2}$$

Proof. We show that $D_1 f(x)$ exists almost everywhere, the argument for $D_i f$ being identical. Write $x = (x_1, x_2, \dots, x_n)$ as $x = (x_1, x')$ where $x' = (x_2, \dots, x_n)$. Then for any $x' \in \mathbf{R}^{n-1}$ let

$$N_{x'} := \{x_1 \in \mathbf{R} : D_1 f(x_1, x') \text{ does not exist.}\}.$$

Then by the one variable result $N_{x'}$ is a set of measure zero in $\mathbf{R}$ for all $x' \in \mathbf{R}^{n-1}$. Therefore by Fubini's the set

$$N = \bigcup_{x' \in \mathbf{R}^{n-1}} N_{x'} = \{x \in \mathbf{R}^n : D_1 f(x) \text{ does not exist}\}$$

is a set of measure zero. That $|D_i f(x)| \le M$ at points where it exists is clear (or follows from the one dimensional result). At points where $Df(x)$ exists

$$\|Df(x)\| = \sqrt{D_1 f(x)^2 + \cdots + D_n f(x)^2} \le \sqrt{M^2 + \cdots + M^2} = \sqrt{n}M.$$

Finally we show (2.2) in the case of $i = 1$. Using the notation above, Fubini's theorem, and the one variable integration by parts formula.

$$\begin{aligned}
\int_{\mathbf{R}^n} D_1 f(x)\varphi(x)\,dx &= \int_{\mathbf{R}^{n-1}} \int_{\mathbf{R}} D_1 f(x_1, x')\varphi(x_1, x')\,dx_1\,dx' \\
&= -\int_{\mathbf{R}^{n-1}} \int_{\mathbf{R}} f(x_1, x') D_1\varphi(x_1, x')\,dx_1\,dx' \\
&= -\int_{\mathbf{R}^n} f(x) D_1\varphi(x)\,dx
\end{aligned}$$

This completes the proof. □

2.3. Definition. Let $f\colon \mathbf{R}^n \to \mathbf{R}$, then for a fixed vector $v \in \mathbf{R}^n$ define

$$df(x, v) := \lim_{t \to 0} \frac{f(x + tv) - f(x)}{t}.$$

When this limit exists it is ***directional derivative*** of f in the direction of v at the point x. □

2.4. Proposition. *Let $f\colon \mathbf{R}^n \to \mathbf{R}$ be Lipschitz and let $v \in \mathbf{R}^n$ be a fixed vector. Then $df(x, v)$ exists for almost all $x \in \mathbf{R}^n$ and is given by the formula*

$$df(x, v) = Df(x) \cdot v \tag{2.3}$$

for almost all x.

Proof. Note if $v = e_1$ where $e_1, \dots, e_n$ is the standard coordinate basis of $\mathbf{R}^n$ then $df(x,v) = df(x,e_1) = D_1 f(x)$ and the fact that $df(x,v)$ exists almost everywhere follows from Proposition 2.2. In the general case if $v \neq 0$ (and the case $v = 0$ is trivial) there is a linear coordinate system $\xi_1, \dots, \xi_n$ on $\mathbf{R}^n$ so that $df(x,v) = \frac{\partial f}{\partial \xi_1}$. But again Proposition 2.2 can be used to see that $df(x,v)$ exits for almost all $x \in \mathbf{R}^n$.

To see that the formula (2.3) holds let $\varphi \in C_0^\infty(\mathbf{R}^n)$. Then as φ is smooth the usual form of the chain rule implies $d\varphi(x,v) = D\varphi(x) \cdot v$. Let M be the Lipschitz constant of f. Then

$$\left| \frac{f(x+tv) - f(x)}{t} \varphi(x) \right| \leq \frac{M\|tv\|}{|t|} |\varphi(x)| \leq M\|v\| \|\varphi\|_{L^\infty}.$$

Therefore for $0 < |t| \leq 1$ the function $x \mapsto \left| \frac{f(x+tv)-f(x)}{t} \varphi(x) \right|$ is uniformly bounded and has compact support. Thus by the dominated convergence theorem and the version of integration by parts given in Proposition 2.2

$$\begin{aligned}
\int_{\mathbf{R}^n} df(x,v)\varphi(x)\,dx &= \lim_{t\to 0} \int_{\mathbf{R}^n} \frac{f(x+tv)-f(x)}{t}\varphi(x)\,dx \\
&= \lim_{t\to 0} \frac{1}{t}\left(\int_{\mathbf{R}^n} f(x+tv)\varphi(x)\,dx - \int_{\mathbf{R}^n} f(x)\varphi(x)\,dx \right) \\
&= \lim_{t\to 0} \frac{1}{t}\left(\int_{\mathbf{R}^n} f(x)\varphi(x-tv)\,dx - \int_{\mathbf{R}^n} f(x)\varphi(x)\,dx \right) \\
&= \lim_{t\to 0} \int_{\mathbf{R}^n} f(x)\frac{\varphi(x-tv)-\varphi(x)}{t}\,dx \\
&= \int_{\mathbf{R}^n} f(x) d\varphi(x,-v)\,dx \\
&= -\int_{\mathbf{R}^n} f(x) D\varphi(x)\cdot v\,dx \\
&= -\sum_{i=1}^n v_i \int_{\mathbf{R}^n} f(x) D_i\varphi(x)\,dx \\
&= \sum_{i=1}^n v_i \int_{\mathbf{R}^n} D_i f(x)\varphi(x)\,dx \\
&= \int_{\mathbf{R}^n} Df(x)\cdot v\varphi(x)\,dx.
\end{aligned}$$

Thus $\int_{\mathbf{R}^n} df(x,v)\varphi(x)\,dx = \int_{\mathbf{R}^n} Df(x)\cdot v\varphi(x)\,dx$ for all $\varphi \in C^\infty(\mathbf{R}^n)$. Therefore $d(x,v) = Df(x)\cdot v$ for almost all $x \in \mathbf{R}^n$. □

2.5. Definition. Let $f\colon \mathbf{R}^n \to \mathbf{R}^m$. Then f is ***differentiable*** at $x_0 \in \mathbf{R}^n$ iff there is a linear map $L\colon \mathbf{R}^n \to \mathbf{R}^m$ so that

$$f(x) - f(x_0) = L(x - x_0) + o(\|x - x_0\|).$$

In this case the linear map L is easily seen to be unique and will be denoted by $f'(x_0)$. □

By $o(\|x - x_0\|)$ we mean a function of the form $\|x - x_0\| g(x, x_0)$ where $\lim_{x \to x_0} g(x, x_0) = 0$. This definition could be given a little more formally by letting $S^{n-1} = \{u \in \mathbf{R}^n : \|u\|\}$ be the unit sphere in $\mathbf{R}^n$. Then $f \colon \mathbf{R}^n \to \mathbf{R}^m$ is differentiable at x with $f'(x) = L$ iff for all $\varepsilon > 0$ there is a $\delta > 0$ so that for all $u \in S^{n-1}$

$$0 < |t| \leq \delta \quad \text{implies} \quad \left\| \frac{f(x_0 + tu) - f(x_0)}{t} - Lu \right\| < \varepsilon. \tag{2.4}$$

2.6. Theorem (Rademacher [6] 1919)**.** *If $f \colon \mathbf{R}^n \to \mathbf{R}^m$ is Lipschitz, say $\|f(x_1) - f(x_0)\| \leq M\|x_1 - x_0\|$, then the derivative $f'(x)$ exists for almost all $x \in \mathbf{R}^n$. In the case $m = 1$, so that $f \colon \mathbf{R}^n \to \mathbf{R}$, then $f'(x)$ is given by*

$$f'(x)v = Df(x) \cdot v$$

for almost all $x \in \mathbf{R}^n$.

Proof. We first consider the case when $m = 1$ so that f is scalar valued. Let E_0 be the set of points $x \in \mathbf{R}^n$ so that $Df(x)$ exists. By Proposition 2.2 the set E_0 has full measure (that is $\mathcal{L}^n(\mathbf{R}^n \setminus E_0) = 0$). Let $\mathbf{R}^{\#} = \mathbf{R} \setminus \{0\}$ and define $Q \colon E_0 \times S^{n-1} \times \mathbf{R}^{\#} \to [0, \infty)$ to be

$$Q(x, u, t) := \left| \frac{f(x + tu) - f(x)}{t} - Df(x) \cdot u \right|.$$

Then, in light of the form of the definition of differentiable given by 2.4, we wish to show that $Q(x, u, t)$ can be made small by making t small.

First note if $u, u' \in S^{n-1}$ then, using the Lipschitz condition of f and the bound (2.1) on $\|Df(x)\|$ we have

$$\begin{aligned} |Q(x, u, t) - Q(x, u', t)| &\leq \left| \frac{f(x + tu) - f(x + tu')}{t} \right| + \left| Df(x) \cdot (u - u') \right| \\ &\leq \frac{M|t|\|u - u'\|}{|t|} + \|Df(x)\|\|u - u'\| \\ &\leq M(1 + \sqrt{n})\|u - u'\|. \end{aligned} \tag{2.5}$$

Now choose a sequence $\{u_k\}_{k=1}^{\infty}$ that is dense in S^{n-1}. For each $k = 1, 2, \ldots$ set

$$E_k = \{x \in E_0 : df(x, u_k) \text{ exists and } df(x, u_k) = Df(x) \cdot u_k\}.$$

Then by Proposition 2.4 each set E_k has full measure and thus the same is true of

$$E := \bigcap_{k=1}^{\infty} E_k.$$

Fix a point $x \in E$. Let $\varepsilon > 0$. Then there is a K so that $\{u_k\}_{k=1}^K$ is $\varepsilon/(2M(1+\sqrt{n})$ dense in S^{n-1}. That is for all $u \in S^{n-1}$ there is a $k \leq K$ so that

$$M(1+\sqrt{n})\|u-u_k\| < \frac{\varepsilon}{2}. \tag{2.6}$$

As each directional derivative $df(x,u_k)$ exists there is a $\delta > 0$ so that if $1 \leq k \leq K$ then

$$0 < |t| \leq \delta \quad \text{implies} \quad Q(x,u_k,t) < \frac{\varepsilon}{2}.$$

Then for any $u \in S^{n-1}$ there is a u_k with $k \leq K$ so that (2.6) holds. Therefore $0 < |t| \leq \delta$ implies

$$\begin{aligned} Q(x,u,t) &\leq Q(x,u_k,t) + |Q(x,u,t) - Q(x,u_k,t)| \\ &< \frac{\varepsilon}{2} + M(1+\sqrt{n})\|u-u_k\| \\ &< \frac{\varepsilon}{2} + \frac{\varepsilon}{2} = \varepsilon. \end{aligned}$$

This shows that f is differentiable at x with $f'(x)v = Df(x)\cdot v$. As x was any point of E and E has full measure this completes the proof in the case $m = 1$.

For $m \geq 2$ write f as

$$f(x_1,\dots,x_n) = f(x) = \begin{bmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_m(x) \end{bmatrix}$$

Let $F_i = \{x : f_i'(x) \text{ exists and } f_i'(x)v = Df_i(x)\cdot v\}$. Then we have just shown that F_i has full measure and thus the same is true of $F := E_1 \cap \cdots \cap F_m$. For $x \in F$ let $L_x \colon \mathbf{R}^n \to \mathbf{R}^m$ be the linear map given by the matrix

$$L_x := \begin{bmatrix} D_1 f_1(x) & D_2 f_1(x) & \cdots & D_n f_1(x) \\ D_1 f_2(x) & D_2 f_2(x) & \cdots & D_n f_2(x) \\ \vdots & \vdots & \ddots & \vdots \\ D_1 f_m(x) & D_2 f_m(x) & \cdots & D_n f_m(x) \end{bmatrix}.$$

Then using that the derivatives $f_i'(x)$ for all $x \in F$ it is not hard to show that $f'(x)$ exists for all $x \in F$ and that for these x the derivative is given by $f'(x) = L_x$. This competes the proof. □

3. A General Sard Type Theorem for Maps Between Spaces of the Same Dimension

Let $G \colon \mathbf{R}^n \to \mathbf{R}^n$. Then define the set of ***critical points*** of G as

$$\operatorname{Crit}(G) := \{x : G'(x) \text{ does not exist or } G'(x) \text{ exists but } \det G'(x) = 0\}.$$

The set of ***regular points*** of G is $\mathbf{R}^n \setminus \operatorname{Crit}(G)$. Thus x is a regular value iff $G'(x)$ exists and is a linear automorphism of $\mathbf{R}^n$. The set of ***critical values*** of G is $G[\mathrm{Crit}(G)]$ and the set of ***regular values*** is $\mathbf{R}^n \setminus G[\mathrm{Crit}(G)]$. Thus y is a regular value iff when ever there is an $x \in \mathbf{R}^n$ with $G'(x) = y$ then $G'(x)$ exists and is nonsingular. (Note that if y is not in the image of G then will be regular value of G. Thus non-values of G are regular values.)

3.1. Theorem. *Let $G\colon \mathbf{R}^n \to \mathbf{R}^n$ be Lipschitz. Then the set $G[\mathrm{Crit}(G)]$ of critical values of G has measure zero (that is $\mathcal{L}^n G[\mathrm{Crit}(G)] = 0$). (Or what is the same thing almost very point $y \in \mathbf{R}^n$ is a regular value of G.)*

This will be based on some more general results. We will denote the Lebesgue outer measure on $\mathbf{R}^n$ by $\mathcal{L}^n$. (That is we use the same notation for the outer measure and the measure. As these agree on the set of measurable functions this should not lead to any confusion.)

3.2. Proposition. *Let $f\colon \mathbf{R}^n \to \mathbf{R}^n$ be an arbitrary map (it need not be continuous or even measurable) and assume that there is a set A (which need not be measurable) so that at every point $a \in A$ the derivative $f'(a)$ exists and satisfies $|\det f'(a)| \le M$ for some constant M. Then the outer measure of $f[A]$ satisfies $\mathcal{L}^n(f[A]) \le M\mathcal{L}^n(A)$. In particular if A has measure zero then so does $f[A]$.*

The proof will be based on a weaker version of the result (which is really all that is needed to prove Theorem 3.1).

3.3. Lemma. *There is a constant $C(n)$ only depending on the dimension so the if $f\colon \mathbf{R}^n \to \mathbf{R}^n$ is an arbitrary map (it need not be continuous or even measurable) and A is a set (which need not be measurable) so that at every point $a \in A$ the derivative $f'(a)$ exists and satisfies $|\det f'(a)| \le M$ for some constant M. Then the outer measure of $f[A]$ satisfies $\mathcal{L}^n(f[A]) \le C(n)M\mathcal{L}^n(A)$.*

Proof. Let $a \in A$ and let $\alpha(x) := f(a) + f'(a)(x-a)$ be the affine approximation to f at a. Then if $B(a,r)$ is the ball of radius r centered at a then from the change of variable formula from calculus we have $\mathcal{L}^n(\alpha[B]) = |\det f'(a)|\mathcal{L}^n(B(a,r))$.

3.4. Claim. If $a \in A$ and $\varepsilon > 0$ then there is an an $r_0 = r_0(x_0, \varepsilon)$ so that if $r \le r_0(x_0, \varepsilon)$ then

$$\mathcal{L}^n(f[B(a,r)]) \le (M+\varepsilon)\mathcal{L}^n(B(a,r)) \tag{3.1}$$

(The proof will show this also holds for closed balls $\overline{B}(a,r)$ with $r \le r_0$.)

Proof of the claim. As f is differentiable at a there is a monotone decreasing function $\omega\colon [0,\infty) \to [0,\infty)$ with $\lim_{t\searrow 0} \omega(t) = 0$ so that

$$\|f(x) - f(a) - f'(a)(x-a)\| \le \omega(\|x-a\|)\|x-a\|.$$

(To be explicit we can take $\omega(t) = \sup\{\|f(x) - f(a) - a'(a)(x-a)\|/\|x-a\| : 0 < \|x-a\| \le t\}$.) Let $\alpha(x)\colon \mathbf{R}^n \to \mathbf{R}^n$ be the affine map $\alpha(x) =$

$f(a) + f'(a)(x - a)$. Then $\|f(x) - \alpha(x)\| \leq \omega(\|x - a\|)\|x - a\|$. As above $\mathcal{L}^n(\alpha[B(a, r)]) = |\det f'(a)|\mathcal{L}^n(B(x, r))$. But if $\|x - a\| \leq r$ then

$$\|x - a\| \leq r \quad \text{implies} \quad \|f(x) - \alpha(x)\| \leq \omega(r)r. \tag{3.2}$$

Let $\overline{B}^n$ be the unit ball in $\mathbf{R}^n$. Then for $\rho > 0$ the set $\rho\overline{B}^n = \{\rho x : x \in \overline{B}^n\}$ is the closed ball of radius ρ and for any set $C \subset \mathbf{R}^n$ the set $C + \rho\overline{B}^n = \{c + \rho x : c \in C, x \in \overline{B}^n\}$ is the tube of radius ρ about C (that is the set of points at a distance $\leq \rho$ from C). But then (3.2) implies

$$f[B(a, r)] \subseteq \alpha[B(a, r)] + \omega(r)r\overline{B}^n.$$

Thus

$$\begin{aligned} \mathcal{L}^n(f[B(a, r)]) &\leq \mathcal{L}^n(\alpha[B(a, r)] + \omega(r)r\overline{B}^n) \\ &= \mathcal{L}^n(\alpha[B(a, 1)] + \omega(r)\overline{B}^n)r^n \end{aligned}$$

But $\lim_{t\searrow 0} \omega(t) = 0$ implies

$$\begin{aligned} \lim_{r\searrow 0} \mathcal{L}^n(\alpha[B(a, 1)] + \omega(r)\overline{B}^n) &= \mathcal{L}^n(\alpha[B(a, 1)] \\ &= |\det f'(a)|\mathcal{L}^n(B(a, 1)) \\ &\leq M\mathcal{L}^n(B(a, 1)) \end{aligned}$$

Thus there is an r_0 so that $\mathcal{L}^n(\alpha[B(a, 1)] + \omega(r_0)\overline{B}^n) \leq (M + \varepsilon)\mathcal{L}^n(B(a, 1))$. Then for $r \leq r_0$ our bounds imply

$$\begin{aligned} \mathcal{L}^n(f[B(a, r)]) &\leq \mathcal{L}^n(\alpha[B(a, 1)] + \omega(r)\overline{B}^n)r^n \\ &\leq (M + \varepsilon)\mathcal{L}^n(B(a, 1))r^n = (M + \varepsilon)\mathcal{L}^n(B(a, r)) \end{aligned}$$

which shows (3.1) holds and completes the proof of the claim. □

Before returning to the proof of Theorem 3.6 we need a covering theorem from analysis. This is the Besicovitch covering theorem which in many ways is more useful than the Vitali covering theorem. There are many equivalent statements of the theorem and at first glance the following many not look like the form given in some texts. For a proof see [8, Thm 1.3.5 p. 9].

3.5. Theorem (Besicovitch [2, 1945])**.** *There is a number $N = N(n)$ so that if $A \subseteq \mathbf{R}^n$ and $r\colon A \to (0, \infty)$ with $\sup_{a\in A} r(a) < \infty$ then there is a subset $\{a_k\}_{k=1}^{\infty} \subseteq A$ so that*

$$A \subset \bigcup_{k=1}^{\infty} \overline{B}(a_k, r(a_k))$$

and for all $x \in \mathbf{R}^n$

$$\#\{k : x \in \overline{B}(a_k, r(a_k))\} \leq N(n).$$

(That is any point of $\mathbf{R}^n$ is in at most $N(n)$ of the balls $\overline{B}(a_k, r(a_k))$.) □

We now return to the proof of Lemma 3.3. If $\mathcal{L}^n(A) = \infty$ there is nothing to prove. Thus assume $\mathcal{L}^n(A) < \infty$. Then there is an open set U so that $A \subset U$ and $\mathcal{L}^n(U) \leq 2\mathcal{L}^n(A)$. Fix $\varepsilon > 0$. For each $x \in A$ we can use the claim to find an $r(x) > 0$ so that the closed ball $\overline{B}(x, r(x)) \subset U$ and

$$\mathcal{L}^n f[\overline{B}(x, r(x))] \leq (M + \varepsilon)\mathcal{L}^n B(x, r(x)).$$

As $\overline{B}(x, r(x)) \subset U$ and $\mathcal{L}^n(U) \leq 2\mathcal{L}^n(A) < \infty$ we have $\sup_{x\in A} r(x) < \infty$. Therefore by the Besicovitch theorem there is a subset $\{x_k\}_{k=1}^{\infty} \subset A$ so that if $r_k := r(x_k)$ then

$$A \subseteq \bigcup_{k=1}^{\infty} \overline{B}(x_k, r_k)$$

and each $x \in \mathbf{R}^n$ is in at most $N(n)$ of the balls $B(x_k, r_k)$. This last fact, along with the fact that each $\overline{B}(x_k, r_k) \subset U$, implies $\sum_{k=1}^{\infty} \chi_{\overline{B}(x_k,r_k)}(x) \leq N(n)\chi_U(x)$ (where χ_S is the characteristic function of the set S). Therefore

$$\begin{aligned}
\sum_{k=1}^{\infty} \mathcal{L}^n(\overline{B}(x_k, r_k)) &= \int_{\mathbf{R}^n} \sum_{k=1}^{\infty} \chi_{\overline{B}(x_k,r_k)}(x)\, dx \\
&\leq \int_{\mathbf{R}^n} N(n)\chi_U(x)\, dx \\
&= N(n)\mathcal{L}^n(U) \\
&\leq 2N(n)\mathcal{L}^n(A).
\end{aligned}$$

We now estimate $\mathcal{L}^n(f[A])$.

$$\begin{aligned}
\mathcal{L}^n(f[A]) &\leq \mathcal{L}^n\Big(f\Big[\bigcup_{k=1}^{\infty} B(x_k, r_k)\Big]\Big) \\
&\leq \sum_{k=1}^{\infty} \mathcal{L}^n(f[B(x_k, r_k)]) \\
&\leq (M + \varepsilon) \sum_{k=1}^{\infty} \mathcal{L}^n(B(x_k, r_k)) \\
&\leq 2N(n)(M + \varepsilon)\mathcal{L}^n(A)
\end{aligned}$$

As ε was arbitrary this implies $\mathcal{L}^n(f[A]) \leq C(n)\mathcal{L}^n(A)$ where $C(n) = 2N(n)$. This completes the proof. □

3.6. Proposition. *Let $f : \mathbf{R}^n \to \mathbf{R}^n$ be a map (which is not assumed to be continuous or even measurable) and let*

$$\mathcal{C} := \{x \in \mathbf{R}^n : f'(x) \text{ exists, but } \det f'(x) = 0\}$$

be the set of points where $f'(x)$ exists but has rank less than n. Then the set $f[\mathcal{C}]$ has measure zero.

Proof. If $\mathcal{L}^n(\mathcal{C}) < \infty$ then the result follows from Lemma 3.3 by letting $M = 0$. In the general case we can let decompose $\mathcal{C}$ as $\mathcal{C} = \bigcup_{k=1}^{\infty} A_k$ where for each k $\mathcal{L}^n(A_k) < \infty$ (for example let $A_k = B(0,k) \cap A$). Then Lemma 3.3 implies $f[A_k]$ has measure zero and thus the same is true of $f[\mathcal{C}] = \bigcup_{k=1}^{\infty} f[A_k]$. □

Proof of Theorem 3.1. Split $\operatorname{Crit}(G)$ into two sets, the first being

$$\mathcal{N} := \{x \in \mathbf{R}^n : G'(x) \text{ does not exist}\}$$

the points where the derivative does not exist and the second being

$$\mathcal{C} := \{x \in \mathbf{R}^n : \det G'(x) = 0\}$$

being the points where $G'(x)$ exists but has rank less than n.

From Rademacher's Theorem 2.6 the set $\mathcal{N}$ has measure zero. But as G is Lipschitz this implies $\mathcal{L}^n(G[\mathcal{N}]) = 0$. But Theorem 3.6 implies $\mathcal{L}^n(G[\mathcal{C}]) = 0$. Thus

$$\mathcal{L}^n(G[\operatorname{Crit}(G)]) \leq \mathcal{L}^n(G[\mathcal{N}]) + \mathcal{L}^n(G[\mathcal{C}]) = 0 + 0 = 0.$$

This completes the proof. □

Proof of Proposition 3.2. If $\mathcal{L}^n(A) = \infty$ there is nothing to prove so assume $\mathcal{L}^n(A) < \infty$. Let $\varepsilon > 0$. Then there is an open set U so that $A \subset U$ and $\mathcal{L}^n(U) \leq \mathcal{L}^n(A) + \varepsilon$.

For each $a \in A$ the claim 3.4 gives an $r_0 = r_0(a, \varepsilon)$ so that for all $r \leq r_0$ the ball $B(a, r)$ satisfies $\mathcal{L}^n(f[B(a,r)]) \leq (M + \varepsilon)\mathcal{L}^n(B(a,r))$. Let

$$\mathcal{V}(a) := \{B(a,r) : r \leq r_0(a, \varepsilon) \text{ and } B(a,r) \subset U\}$$

and set

$$\mathcal{V} : \bigcup_{a \in A} \mathcal{V}(a).$$

Then $\mathcal{V}$ is a fine cover of A (that is every point of A is contained in balls of arbitrarily small balls contained in $\mathcal{V}$. Then by the Vitali covering theorem (cf. [8, Thm 1.3.6 P. 12] where this is deduced from the Besicovitch covering theorem) there is a countable set $\{B(x_k, r_k)\}_{k=1}^{\infty} \subset \mathcal{V}$ so that

$$\mathcal{L}^n\left(A \setminus \bigcup_{k=1}^{\infty} B(x_k, r_k)\right) = 0$$

and $B(x_i, r_i) \cap B(x_j, r_j) = \varnothing$ if $i \neq j$. Let $A_0 = A \setminus \bigcup_{k=1}^{\infty} B(x_k, r_k)$ and $A_1 = A \cap (\bigcup_{k=1}^{\infty} B(x_k, r_k))$. Then $\mathcal{L}^n(A_0) = 0$ and therefore $\mathcal{L}^n(f[A_0]) = 0$

by Lemma 3.3. As to A_1 we have

$$
\begin{aligned}
\mathcal{L}^n(f[A_1]) &\leq \mathcal{L}^n\left(f\left[\bigcup_{k=1}^{\infty} B(x_k, r_k)\right]\right) \\
&\leq \sum_{k=1}^{\infty} \mathcal{L}^n(f[B(x_k, r_k)]) \\
&\leq (M+\varepsilon) \sum_{k=1}^{\infty} \mathcal{L}^n(B(x_k, r_k)) \\
&= (M+\varepsilon)\mathcal{L}^n\left(\bigcup_{k=1}^{\infty} B(x_k, r_k)\right) \\
&\leq (M+\varepsilon)\mathcal{L}^n(U) \\
&\leq (M+\varepsilon)(\mathcal{L}^n(A)+\varepsilon).
\end{aligned}
$$

But ε was arbitrary so $\mathcal{L}^n(f[A_1]) \leq M\mathcal{L}^n(A_1)$. Thus $\mathcal{L}^n(f[A]) \leq \mathcal{L}^n(f[A_0]) + \mathcal{L}^n(f[A_1]) \leq 0 + M\mathcal{L}^n(A_1) = \mathcal{L}^n(A)$ as $\mathcal{L}^n(A_0) = 0$. This completes the proof. □

3.1. **A more general result.**

3.7. Theorem. *Let $f\colon \mathbf{R}^n \to \mathbf{R}^n$ be any map and let $A \subset \mathbf{R}^n$ be a set so that $f'(x)$ exists for all $x \in A$. Then*

$$
\int_{\mathbf{R}^n} \#(A \cap f^{-1}[y])\, dy \leq \int_A |\det f'(x)|\, dx.
$$

Note that $\chi_{f[A]}(y) \leq \#(A \cap f^{-1}[y])$ for all y so there is the inequality $\mathcal{L}^n(f[A]) \leq \int_{\mathbf{R}^n} \#(A \cap f^{-1}[y])\, dy$ therefore the theorem as a corollary (which will actually be proven first as for use as a lemma):

3.8. Corollary. *Let $f\colon \mathbf{R}^n \to \mathbf{R}^n$ be any map and let $A \subseteq \mathbf{R}^n$ be a set so that $f'(x)$ exists for all $x \in A$. Then*

$$
\mathcal{L}^n(f[A]) \leq \int_A |\det f'(x)|\, dx.
$$

Proof. We can assume that $\mathcal{L}^n(A) < \infty$ for if we prove the result in that case we can split a set A with $\mathcal{L}^n(A) = \infty$ into a countable disjoint collection of subsets with finite measure and apply what we have shown to those subsets and then sum over the collection to get the bound in the general case. So assume $\mathcal{L}^n(A) < \infty$ and define

$$\begin{aligned}
A_0 &:= \{x \in A : \det f'(x) = 0\} \\
A_{k,j} &:= \left\{ x \in A : \frac{j-1}{2^k} < |\det f'(x)| \le \frac{j}{2^k} \right\} \\
A_k &:= \bigcup_{j=1}^{2^{2k}} A_{k,j} \\
f_k &:= \sum_{j=1}^{2^{2k}} \frac{j-1}{2^k} \chi_{A_{k,j}}
\end{aligned}$$

where $\chi_{A_{k,j}}$ is the characteristic function of $A_{k,j}$. Then $0 \le f_k \le f_{k+1}$, $\lim_{k\to\infty} f_k(x) = |\det f'(x)|$ on A, and the support of f_k is A_k. By the monotone convergence theorem

$$\lim_{k\to\infty} \int_A f_k(x)\, dx = \int_A |\det f'(x)|\, dx.$$

By Proposition 3.2

$$\mathcal{L}^n(f[A_{k,j}]) \le \frac{j}{2^k} \mathcal{L}^n(A_{k,j}).$$

Thus

$$\begin{aligned}
\mathcal{L}^n(f[A_k]) &\le \sum_{j=1}^{2^{2k}} \mathcal{L}^n(f[A_{k,j}]) \\
&\le \sum_{j=1}^{2^{2k}} \frac{j}{2^k} \mathcal{L}^n(A_{k,j}) \\
&= \frac{1}{2^k} \sum_{j=1}^{2^{2k}} \mathcal{L}^n(A_{k,j}) + \sum_{j=1}^{2^{2k}} \frac{j-1}{2^k} \mathcal{L}^n[A_{k,j}] \\
&= \frac{1}{2^k} \mathcal{L}^n(A_k) + \sum_{j=1}^{2^{2k}} \frac{j-1}{2^k} \mathcal{L}^n[A_{k,j}] \\
&= \frac{1}{2^k} \mathcal{L}^n(A_k) + \int_A f_k(x)\, dx \\
&\le \frac{1}{2^k} \mathcal{L}^n(A) + \int_A f_k(x)\, dx
\end{aligned}$$

Also $f[A_k] \subseteq f[A_{k+1}]$ and $\bigcup_{k=1}^{\infty} f[A_k] = f[A]$ so

$$\begin{aligned}\mathcal{L}^n(f[A]) = \lim_{k\to\infty} \mathcal{L}^n(f[A_k])& \\ &\le \lim_{k\to\infty}\left(\frac{1}{2^k}\mathcal{L}^n(A_k) + \int_A f_k(x)\,dx\right) \\ &= \int_A |\det f'(x)|\,dx.\end{aligned}$$

This completes the proof. □

proof of Theorem 3.7. We can assume that $\det f'(x) \neq 0$ for all $y \in A$. This is because if $N := \{x \in A : \det f'(x) = 0\}$ then $\int_A |\det f'(x)|\,dx = \int_{A\setminus N} |\det f'(x)|\,dx$ and by Proposition 3.6 the image $f[N]$ has measure zero and thus $\int_{\mathbf{R}^n} \#(A \cap f^{-1}[y])\,dy = \int_{\mathbf{R}^n\setminus f[N]} \#(A\cap f^{-1}[y])\,dy$.

As we are assuming that $f'(a)$ is nonsingular for all $a \in A$ the inverse $f'(a)$ exists. Define

$$\begin{aligned}C_k = \Big\{ a \in A :& \|f'(a)\|_{\mathrm{Op}}, \|f'(a)\|_{\mathrm{Op}} \le k, \text{ and } \|f(x) - f(a) - f'(a)(x-a)\| \\ &\le \frac{1}{\|f'(a)^{-1}\|_{\mathrm{Op}}\|x-a\|} \text{ for } \|x-a\| \le \frac{1}{k}\Big\}\end{aligned}$$

Then $A = \bigcup_{k=1}^{\infty} C_k$.

3.9. Claim. If $a \in A_k$ and if $a \in \mathbf{R}^n$ with $\|x-a\| \le \frac{1}{k}$ then

$$\frac{1}{2k}\|x-a\| \le \|f(x)-f(a)\| \le \frac{3}{2}k\|x-a\|.$$

Proof of claim. Note that $\|f'(a)\|_{\mathrm{Op}} \le k$ implies

$$1 = \|I\|_{\mathrm{Op}} = \|f'(a)f'(a)^{-1}\|_{\mathrm{Op}}\| \le \|f'(a)\|_{\mathrm{Op}}\|f'(a)^{-1}\|_{\mathrm{Op}} \le k\|f'(a)\|_{\mathrm{Op}}\|f'(a)^{-1}\|_{\mathrm{Op}}$$

so that $\frac{1}{\|f'(a)^{-1}\|_{\mathrm{Op}}} \le k$. Thus

$$\begin{aligned}\|f(x)-f(a)\| &\le \|f(x)-f(a)-f'(a)(x-a)\| + \|f'(a)(x-a)\| \\ &\le \frac{1}{2\|f'(a)^{-1}\|_{\mathrm{Op}}}\|x-a\| + \|f'(a)\|_{\mathrm{Op}}\|x-a\| \\ &\le \frac{k}{2}\|x-a\| + k\|x-a\| \\ &= \frac{3}{2}k\|x-a\|\end{aligned}$$

which is the required upper bound on $\|f(x) - f(a)\|$. For the lower bound estimate

$$\begin{aligned}\|x-a\| &\le \|f'(a)^{-1}\|_{\mathrm{Op}}\|f'(a)(x-a)\| \\ &\le \|f'(a)^{-1}\|\big(\|f(x)-f(a)-f'(a)(x-a)\| + \|f(x)-f(a)\|\big) \\ &\le \|f'(a)^{-1}\|_{\mathrm{Op}}\frac{1}{2\|f'(a)^{-1}\|_{\mathrm{Op}}}\|x-a\| + \|f'(a)^{-1}\|_{\mathrm{Op}}\|f(x)-f(a)\| \\ &\le \frac{1}{2}\|x-a\| + k\|f(x)-f(a)\|.\end{aligned}$$

This can be solved for $\|f(x) - f(a)\|$ to get

$$\frac{1}{2k}\|x-a\| \le \|f(x)-f(a)\|.$$

This completes the proof of the claim. $\square$

Returning to the proof of Theorem 3.7 we let $A_1 := C_1$ and $A_k := C_k \setminus C_{k-1}$ for $k \ge 2$. Then we can decompose each A_k into a disjoint union

$$A_k = \bigcup_{j=1}^{\infty}\!\!\!\cdot\;\, A_{k,j} \quad \text{with} \quad \operatorname{diameter}(A_{k,j}) \le \frac{1}{k}.$$

Then by the claim we have that if $a, b \in A_{k,j}$ then

$$\frac{1}{2k}\|b-a\| \le \|f(b)-f(a)\| \le \frac{3}{2}k\|b-a\|.$$

Therefore the restriction $f\big|_{A_{k,j}} : A_{k,j} \to \mathbf{R}^n$ is injective. Also A is a disjoint union $A = \dot\bigcup_{k,j} A_{k,j}$. Now from Corollary 3.8

$$\mathcal{L}^n(f[A_{k,j}]) \le \int_{A_{k,j}} |\det f'(x)|\,dx.$$

But as $f\big|_{A_{k,j}}$ is injective we have

$$\begin{aligned}\int_{\mathbf{R}^n} \#(A \cap f^{-1}[y])\,dy &= \sum_{j,k}\mathcal{L}^n(f[A_{k,j}]) \\ &\le \sum_{j,k}\int_{A_{k,j}} |\det f'(x)|\,dx \\ &= \int_A |\det f'(x)|\,dx.\end{aligned}$$

This completes the proof. $\square$

4. An Inverse Function Theorem for Continuous Functions Differentiable at a Single Point

Let $A\colon \mathbf{R}^n \to \mathbf{R}^m$ be a linear map. Then the ***operator norm*** of A is defined by

$$\|A\|_{\mathrm{Op}} := \inf_{0\neq v\in\mathbf{R}^n} \frac{\|Av\|}{\|v\|}.$$

Therefore the inequality

$$\|Av\| \le \|A\|_{\mathrm{Op}}\|v\|.$$

holds.

We now give a form of the inverse function theorem that only requires the function $G\colon \mathbf{R}^n \to \mathbf{R}^n$ to be continuous and differentiable at one point x_0. In this generality the function need not have a local inverse, but we will be able to solve the equation $G(x) = y$ for y close to $G(x_0)$. However the proof is just an easy variant on the usual proof of the inverse function theorem where the Brouwer fixed point theorem is used instead of Banach's fixed point theorem for contractions.

4.1. Theorem. *Let $G\colon \mathbf{R}^n \to \mathbf{R}^n$ be continuous and differentiable at x_0 with $G'(x_0)$ nonsingular. Set $y_0 = G(x_0)$. Then there is an $r_0 > 0$ so that if*

$$\beta := \frac{1}{2\|G'(x_0)^{-1}\|_{\mathrm{Op}}}, \qquad r_1 = \beta r_0,$$

then

1. *for all $y \in \overline{B}(y_0, r_1)$ there is an $x \in \overline{B}(x_0, r_0)$ with $G(x) = y$.*
2. *If $y \in \overline{B}(y_0, r_1)$ and $x \in \overline{B}(x_0, r_0)$ with $G(x) = y$ the inequalities*

$$\beta\|x - x_0\| \le \|y - y_0\| \le (\beta + \|G'(x_0)\|_{\mathrm{Op}})\|x - x_0\|.$$

hold.

Proof. From the definition of the derivative there is an $r_0 > 0$ so that

$$\begin{aligned} &\|x - x_0\| \le r_0 \quad \text{implies} \\ &\|G(x) - y_0 - G'(x_0)(x - x_0)\| \le \frac{1}{2\|G'(x_0)^{-1}\|_{\mathrm{Op}}}\|x - x_0\| \\ &\qquad\qquad = \beta\|x - x_0\|. \end{aligned} \tag{4.1}$$

For $y \in \mathbf{R}^n$ define $\Phi_y\colon \mathbf{R}^n \to \mathbf{R}^n$ by

$$\Phi_y(x) := x - G'(x_0)^{-1}(G(x) - y).$$

Then

$$\Phi_y(x) = x \quad \text{if and only if} \quad G(x) = y \tag{4.2}$$

and

$$\begin{aligned} x_0 - \Phi_y(x) &= x_0 - x + G'(x_0)^{-1}(G(x) - y) \\ &= x_0 - x + G'(x_0)^{-1}(G(x) - y_0) + G'(x_0)^{-1}(y_0 - y) \\ &= G'(x_0)^{-1}(G(x) - y_0 - G'(x_0)(x - x_0)) + G'(x_0)^{-1}(y_0 - y) \end{aligned}$$

Using the definition of r_0 we have that if $\|x - x_0\| \le r_0$ then

$$\begin{aligned} \|x_0 - \Phi_y(x)\| &\le \|G'(x_0)^{-1}\|_{\mathrm{Op}} \|G(x) - y_0 - G'(x_0)(x - x_0)\| \\ &\qquad + \|G'(x_0)^{-1}\|_{\mathrm{Op}} \|y - y_0\| \\ &\le \|G'(x_0)^{-1}\|_{\mathrm{Op}} \frac{1}{2\|G'(x_0)^{-1}\|_{\mathrm{Op}}} \|x - x_0\| \\ &\qquad + \|G'(x_0)^{-1}\|_{\mathrm{Op}} \|y - y_0\| \\ &= \frac{1}{2}\|x - x_0\| + \|G'(x_0)^{-1}\|_{\mathrm{Op}} \|y - y_0\| \end{aligned} \tag{4.3}$$

Therefore using the definition of β and r_1

$$\|x - x_0\| \le r_0, \ \|y - y_0\| \le r_1 \quad \text{implies} \quad \|\Phi_y(x) - x_0\| \le r_0.$$

So if $y \in \overline{B}(y_0, r_1)$ we have that $\Phi_y \colon \overline{B}(x_0, r_0) \to \overline{B}(x_0, r_0)$. Thus by the Brouwer fixed point theorem the map Φ_y will have a fixed point in $\overline{B}(x_0, r_0)$. But $\Phi_y(x) = x$ implies that $G(x) = y$ by (4.2). Thus proves the first of the two conclusions of the theorem.

If $x \in \overline{B}(x_0, r_0)$ and $y \in \overline{B}(y_0, r_1)$. Then by (4.2) $\Phi_y(x) = x$ and so by the estimate (4.3)

$$\|x - x_0\| = \|\Phi_y(x) - x_0\| \le \frac{1}{2}\|x - x_0\| + \|G'(x_0)^{-1}\|_{\mathrm{Op}} \|y - y_0\|$$

which along with the definition of β implies

$$\beta \|x - x_0\| = \frac{1}{2\|G'(x_0)^{-1}\|_{\mathrm{Op}}} \|x - x_0\| \le \|y - y_0\|$$

which proves the lower bound of second conclusion of the theorem. To prove the upper we use the implication (4.1)

$$\begin{aligned} \|y - y_0\| &\le \|G(x) - y_0\| \\ &\le \|G(x) - y_0 - G'(x_0)(x - x_0)\| + \|G'(x_0)(x - x_0)\| \\ &\le \beta\|x - x_0\| + \|G'(x_0)\|_{\mathrm{Op}} \|x - x_0\| \\ &(\beta + \|G'(x_0)\|_{\mathrm{Op}})\|x - x_0\|. \end{aligned}$$

This completes the proof. □

5. Derivatives of Set-Valued Functions and Inverses of Lipschitz Functions

A ***set-valued function*** $F\colon \mathbf{R}^n \to \mathbf{R}^m$ is a function so that for each $x \in \mathbf{R}^n$ the value $F(x)$ is a subset of $\mathbf{R}^n$. We will also refer to such functions as ***multiple valued***. If $F(x) = \varnothing$ then F is said to be ***undefined at*** x and if $F(x) \neq \varnothing$ then F is ***defined at*** x. If $f\colon \mathbf{R}^n \to \mathbf{R}^m$ is an ordinary function it determines a set-valued function by $F(x) = \{f(x)\}$. In this case we will just identify F and f and say that F is ***single valued*** and just write $F(x) = f(x)$. More generally if $F(x) = \{y\}$ is single valued at a point x then we will write $F(x) = y$ and conversely if we write $F(x) = y$ this means that F is single valued at x and that $F(x) = \{y\}$.

5.1. Definition. The set-valued function $F\colon \mathbf{R}^n \to \mathbf{R}^m$ is ***differentiable*** at x_0 iff there is a linear map $L\colon \mathbf{R}^n \to \mathbf{R}^m$ so that for all $\varepsilon > 0$ there is a $\delta > 0$ such that

$$\begin{cases} \|x - x_0\| \le \delta, \quad y_0 \in F(x_0), \quad y \in F(x) \quad \text{implies} \\ \|y - y_0 - L(x - x_0)\| \le \varepsilon \|x - x_0\|. \end{cases} \tag{5.1}$$

□

With what I hope is obvious notation if F is single valued at x_0 (and this will always be the case if F is differentiable at x_0) and $y_0 = F(x_0)$ then F is differentiable at x_0 iff

$$y \in F(x) \quad \text{implies} \quad y - y_0 = L(x - x_0) + o(\|x - x_0\|).$$

5.2. Proposition. *If $F\colon \mathbf{R}^n \to \mathbf{R}^m$ is set-valued then*

1. *If F is single valued then F is differentiable at x_0 is the sense of Definition 5.1 if and only if it is differentiable at x_0 in the usual sense (that is is in Definition 2.5.)*
2. *If F is differentiable at x_0 then F is single valued at x_0 and the linear map L in (5.1) is uniquely determined. We will call this linear map **the derivative** of F at x_0 and write $G'(x_0) = L$.*
3. *If F is differentiable at x_0 and $f\colon \mathbf{R}^n \to \mathbf{R}^m$ is single valued and differentiable at x_0 then the set-valued function $H(x) := f(x) + F(x)$ is differentiable at x_0 and $H'(x_0) = f'(x_0) + F'(x_0)$.*

Proof. The first and last of these are straightforward and left to the reader. For the second let $y_0, y_1 \in F(x_0)$. Then letting $x = x_0$ and $y = y_1$ in (5.1) we get $\|y_1 - y_0\| \le 0$. Thus $y_1 = y_0$ which shows F is single valued at x_0. If L and L_1 are linear maps which work in (5.1) then for any $\varepsilon > 0$ let $\delta > 0$ so that (5.1) holds. Then for $y_0 = F(x_0)$ and $y \in F(x)$ where $\|x - x_0\| \le \delta$

$$\begin{aligned} \|(L - L_1)(x - x_0)\| &= \|L(x - x_0) - L_1(x - x_0)\| \\ &\le \|y - y_0 - L(x - x_0)\| + \|y - y_0 - L_1(x - x_0)\| \\ &\le 2\varepsilon \|x - x_0\| \end{aligned}$$

Thus can be rescaled to show for all $v \in \mathbf{R}^n$ that $\|(L - L_1)v\| \le 2\varepsilon\|v\|$. As ε was arbitrary this implies $L = L_1$ and completes the proof. □

5.3. Theorem. *Let $G\colon \mathbf{R}^n \to \mathbf{R}^n$ be a surjective Lipschitz function and assume that for each $y \in \mathbf{R}^n$ the preimage $G^{-1}[y]$ is connected. Then define a set-valued function $F\colon \mathbf{R}^n \to \mathbf{R}^n$ to be the inverse of G, that is*

$$F(y) := G^{-1}[y] = \{x : F(x) = y\}.$$

Then F is differentiable almost everywhere with $F'(G(x)) = G'(x)^{-1}$.

Proof. Let $x_0 \in \mathbf{R}^n$ be a point where $G'(x_0)$ exists and is nonsingular. Let $y_0 = G(x_0)$. We first show that F is differentiable at y_0 with $F'(y_0) = G'(x_0)^{-1}$. Toward this end we use the version of the inverse function given by 4.1 to get positive numbers r_0 and β so that if $r_1 = \beta r_0$ and $C_0 = (\beta + \|G'(x_0)\|_{\mathrm{Op}})$ then

$$\text{For all } y \in \overline{B}(y_0, r_1) \text{ there is } x \in \overline{B}(x_0, r_0) \text{ with } G(x) = y.$$

and

$$\begin{cases} x \in \overline{B}(x_0, r_0) \text{ and } y \in \overline{B}(y_0, r_1) \text{ with } G(x) = y \text{ implies} \\ \beta\|x - x_0\| \le \|y - y_0\| \le C_0\|x - x_0\|. \end{cases} \tag{5.2}$$

If $y \in B(y_0, r_1)$, so that $\|y - y_0\| < r_1 = \beta r_0$, then for $x \in \overline{B}(x_0, r_0)$ we have $\beta\|x - x_0\| \le \|y - y_0\| < \beta r_0$ and so $x \in B(x_0, r_0)$. Thus implies that if $y \in B(y_0, r_1)$ then $G^{-1}[y] \cap \{x : \|x - x_0\| = r_0\} = \varnothing$. But as $G^{-1}[y] \cap \overline{B}(x_0, r_0) \ne \varnothing$ and $F(y) = G^{-1}[y]$ is connected this yields that $G^{-1}[y] \subset B(x_0, r_0)$ for all $y \in B(y_0, r_1)$. Therefore (5.2) can be improved to

$$\begin{cases} x \in \mathbf{R}^n \text{ and } y \in B(y_0, r_1) \text{ with } x \in F(y) \text{ implies} \\ \beta\|x - x_0\| \le \|y - y_0\| \le C_0\|x - x_0\|. \end{cases} \tag{5.3}$$

As G is differentiable at x_0

$$G(x) - y_0 = G'(x_0)(x - x_0) + o(\|x - x_0\|)$$

which, as $G'(x_0)$ is invertible, is equivalent to

$$x - x_0 = G'(x_0)^{-1}(G(x) - y_0) + o(\|x - x_0\|)$$

Let $y \in B(y_0, r_1)$ then this last equation implies

$$x \in F(y) \quad \text{implies} \quad x - x_0 = G'(x_0)^{-1}(G(x) - y_0) + o(\|x - x_0\|).$$

But by the inequalities in (5.3) for $y \in B(y_0, r_1)$ we have $o(\|x - x_0\|) = o(\|y - y_0\|)$ and thus for $y \in B(y_0, r_1)$

$$x \in F(y) \quad \text{implies} \quad x - x_0 = G'(x_0)^{-1}(y - y_0) + o(\|y - y_0\|).$$

This shows that F is differentiable at y_0 with derivative $G'(x_0)^{-1}$ as claimed.

Let

$$E := \{y \in \mathbf{R}^n : y_0 = G(x) \text{ where } G'(x) \text{ exsits and is nonsingular}\}.$$

We have just shown that F is differentiable at each point of E and that at these points $F'(G(x)) = G'(x)^{-1}$. Theorem 3.1 implies $\mathcal{L}^n(\mathbf{R}^n \setminus E) = 0$. This completes the proof. □

6. Alexandrov's Theorem

Let $f\colon \mathbf{R}^n \to \mathbf{R}$ a function. A vector $b \in \mathbf{R}^n$ is a ***lower support vector*** for f at x_0 iff

$$f(x) - f(x_0) \geq b \cdot (x - x_0) \quad \text{for all} \quad x \in R^n.$$

For each $x_0 \in \mathbf{R}^n$ set

$$\nabla f(x_0) := \{b : b \text{ is a lower suport vector for } f \text{ at } x_0\}.$$

Of course for some functions this maybe empty for many (and possibly all) points x_0. The function f is ***convex*** iff $\nabla f(x)$ is nonempty for all $x \in \mathbf{R}^n$ (this is equivalent to the more usual definition of $f((1-t)x_0+tx_1) \leq f((1-t)x_0) + tf(x_1)$ for $0 \leq t \leq 1$ cf. [7, Thm 1.5.9 p. 29]). If f is convex then it is easy to see that $\nabla f(x)$ is a convex set for all x. Our goal is to give a proof of:

6.1. Theorem (Busemann-Feller [3, 1936] and Alexandrov [1, 1939])**.** *Let $f\colon \mathbf{R}^n \to \mathbf{R}$ be convex. Then the set-valued function ∇f is differentiable almost everywhere.*

6.2. *Remark.* When $n = 1$ this follows from the differentiability almost everywhere of monotone functions of one variable. The case $n = 2$ was proven by H. Busemann and W. Feller [3] in 1936. The general case was settled in 1939 by A. D. Alexandrov [1]. □

We start with the following basic property of convex functions.

6.3. Proposition. *Let $f\colon \mathbf{R}^n \to \mathbf{R}$ be convex. Then the set valved function ∇f is monotone in the sense that*

$$b_0 \in \nabla f(x_0),\ b_1 \in \nabla f(x_1) \quad \textit{implies} \quad (b_1 - b_0) \cdot (x_1 - x_0) \geq 0.$$

We will sometime miss use notation and write this in the more easily read form

$$(\nabla f(x_1) - \nabla f(x_0)) \cdot (x_1 - x_0) \geq 0.$$

Proof. From the definitions of $b_0 \in \nabla f(x_0)$ and $b_1 \in \nabla f(x_1)$ we have

$$f(x_1) - f(x_0) \geq b_0 \cdot (x_1 - x_0)$$
$$f(x_0) - f(x_1) \geq b_1 \cdot (x_0 - x_1).$$

The result follows by adding these. □

Also of use is the following which I have seen used in other parts of geometry (for example in the theory of minimal surfaces).

6.4. Proposition. *Let $f\colon \mathbf{R}^n \to \mathbf{R}$ be convex. Then the set-valued map*

$$F(x) = x + \nabla f(x)$$

is surjective (i.e. for all $y \in \mathbf{R}^n$ then is an $x \in \mathbf{R}^n$ so that $y \in F(x)$) and is non-contractive $\|F(x_1) - F(x_0)\| \geq \|x_1 - x_0\|$. More precisely

$$y_1 \in F(x_1) \textit{ and } y_0 \in F(x_0) \quad \textit{implies} \quad \|y_1 - y_0\| \geq \|x_1 - x_0\|. \tag{6.1}$$

The inverse G of F defined by $G(y) = x$ iff $y \in F(x)$ is single valued and Lipschitz. In fact

$$\|G(x_1) - G(x_0)\| \leq \|x_1 - x_0\|. \tag{6.2}$$

Proof. To see that F is surjective let $y \in \mathbf{R}^n$ and let $\varphi\colon \mathbf{R}^n \to \mathbf{R}$ be the function

$$\varphi(x) := \frac{1}{2}\|x\|^2 + f(x) - x \cdot y.$$

As $h_y(x) := \frac{1}{2}\|x\|^2 - x\cdot y$ is convex the function φ is a sum of convex functions and thus is convex. Also, as $h_y(x) = \frac{1}{2}\|x\|^2 - x \cdot y$ is smooth so that its lower support vectors are given by the classical gradient $\nabla h_y(x) = \{Dh_y\} = x - y$, we have (Proposition 5.2) that

$$\nabla\varphi(x_0) = \nabla h_y(x_0) + \nabla f(x_0) = x_0 - y + \nabla f(x_0) = F(x_0) - y.$$

(Here $F(x_0) - y = \{\eta - y : \eta \in F(x_0)\}$.)

Likewise the function

$$\psi(x) := \frac{1}{4}\|x\|^2 + f(x) - x \cdot y = \varphi(x) - \frac{1}{4}\|x\|^2$$

is convex. Let b a lower support vector to ψ at 0. Then the inequality $\psi(x) - \psi(0) \geq b \cdot (x - 0)$ can be rewritten as

$$\varphi(x) \geq \varphi(0) + \frac{1}{4}\|x\|^2.$$

As φ is continuous this implies there is an x_0 so that φ will have a global minimum at x_0. Then 0 will be a lower support vector for φ at x_0. But $0 \in \nabla\varphi(x_0) = F(x_0) - y$ implies that $y \in F(x_0)$. As y was arbitrary this implies F is surjective as claimed.

Let $y_1 \in F(x_1)$ and $y_0 \in F(x_0)$. Then there are $b_1 \in \nabla f(x_1)$ and $b_0 \in \nabla f(x_0)$ so that $y_1 = x_1 + b_1$ and $y_0 = x_0 + b_0$. Then by proposition 6.3

$$\begin{aligned} (y_1 - y_0) \cdot (x_1 - x_0) &= (x_1 + b_1 - x_0 - b_0) \cdot (x_1 - x_0) \\ &= \|x_1 - x_0\|^2 + (b_1 - b_0) \cdot (x_1 - x_0) \\ &\geq \|x_1 - x_0\|^2. \end{aligned}$$

Therefore by the Cauchy Schwartz inequality

$$\|y_1 - y_0\|\|x_1 - x_0\| \geq (y_1 - y_0) \cdot (x_1 - x_0) \geq \|x_1 - x_0\|^2$$

which implies (6.1). In particular this implies that if $y \in F(x_1)$ and $y \in F(x_0)$ then $\|x_1 - x_0\| \leq \|y - y\| = 0$ and so $x_1 = x_0$. Thus the inverse G of F (defined by $G(y) = \{x : y \in F(x)\}$) is single valued. The inequality (6.2) is then equivalent to (6.1). This completes the proof. □

Proof of Theorem 6.1. Let $f\colon \mathbf{R}^n \to \mathbf{R}$ be convex and let $F(x) = x + \nabla f(x)$. Then by the last proposition F is the inverse a Lipschitz function G. Moreover as each set $\nabla f(x)$ is convex the same is true of $F(x) = x + \nabla f(x)$ and thus $F(x)$ is connected. Therefore Theorem 5.3 implies that F is differentiable almost everywhere. But $\nabla f(x) = F(x) - x$ is differentiable at the same points where F is differentiable and so it is also differentiable almost everywhere. □

6.5. Corollary. *If G is the function used in the proof of Proposition 6.4 then there is a convex function $g\colon \mathbf{R}^n \to \mathbf{R}$ so that $G(y) = \nabla g(y)$.*

Proof. Let $h\colon \mathbf{R}^n \to \mathbf{R}$ be the function

$$h(x) := \frac{1}{2}\|x\|^2 + f(x).$$

Then define g by

$$g(y) = \max_{x\in\mathbf{R}^n}(x \cdot y - h(x)).$$

Note $x \cdot y - h(x) = -(\frac{1}{2}\|x\|^2 + f(x) - x \cdot y) = -\varphi(x)$ with φ as in the proof of Proposition 6.4. In the proof of Proposition 6.4 it has shown that φ always obtains its minimum and therefore g will always obtain its maximum. Moreover given y_0 a point x_0 is where the maximum occurs in the definition of g if and only if

$$0 \in \nabla(x\cdot y_0 - h(x))\big|_{x=x_0} = \nabla(x\cdot y_0 - \frac{1}{2}\|x\|^2 - f(x))\big|_{x=x_0} = y_0 - x_0 - \nabla f(x_0).$$

That is if and only if $y_0 \in x_0 + \nabla f(x_0) = F(x_0)$ (with F as in Proposition 6.4). This is the same as $G(y_0) = x_0$. Thus the definition of g implies

$$h(x) + g(y) \geq x \cdot y \quad \text{with equality iff} \quad x = G(y). \tag{6.3}$$

But this is symmetric in x and y so we also have

$$h(x) = \max_{y\in\mathbf{R}^n}(x \cdot y - g(y)) \tag{6.4}$$

and for a given x the maximum occurs for those y with $G(y) = x$. We now want to show that g is convex. Let $y_0 \in \mathbf{R}^n$ and let $x_0 = G(y_0)$. Then the formula giving h as a maximum implies

$$x_0 \cdot y_0 - g(y_0) \geq x_0 \cdot y - g(y)$$

which can be rewritten as

$$g(y) - g(y_0) \geq x_0 \cdot (y - y_0)$$

and thus x_0 is a lower support vector for g at y_0 and so $\nabla g(y_0) \neq \varnothing$. As y_0 was any point of $\mathbf{R}^n$ this implies g is convex.

Finally if $y_0 \in \mathbf{R}^n$ and x_0 is the point where the maximum occurs in 6.4 then equality holds in (6.3) and so $G(y_0) = x_0$. But also we have $0 \in \nabla(x_0 \cdot y - g(y))\big|_{y=y_0} = x_0 - \nabla g(y_0)$. Thus $x_0 \in \nabla g(y_0)$ if and only if $G(y_0) = x_0$. Therefore $G(y_0) = \nabla g(y_0)$. This completes the proof. □

7. Symmetry of the Second Derivatives

If $f\colon \mathbf{R}^n \to \mathbf{R}$ is convex then at the points x_0 where the derivative of the set-valued map ∇f exists then it is a linear map $(\nabla f)'(x_0)\colon \mathbf{R}^n \to \mathbf{R}^n$. Then we wish to show that at least for almost all x_0 that $\langle(\nabla f)'(x_0)u, v\rangle = \langle(\nabla f)'(x_0)v, u\rangle$. Formally the matrix of $(\nabla f)'$ is $[D_iD_jf]$ and at least on the level of distributions $D_iD_jf = D_jD_if$. The complication is that the distributional second derivatives of f do not have to be functions and so information about them does not seem to directly imply information about $(\nabla f)'(x_0)$ at points where it exists. (For an example in one dimension where the second derivative is not a function consider $f(x) = |x|$. Then $f''(x) = 2\delta$ where δ is the point mass (i.e. Dirac delta function) at the origin.)

7.1. Theorem. *Let $f\colon \mathbf{R}^n \to \mathbf{R}$ be convex. Then for almost all $x \in \mathbf{R}^n$ the derivative of the set-valued function ∇f is single valued and $(\nabla f)'(x)$ is a symmetric positive definite matrix. (Explicitly $\langle(\nabla f)'(x)u, v\rangle = \langle(\nabla f)'(x)v, u\rangle$ and $(\nabla f)'(x)v, v\rangle \geq 0$ for almost all x.)*

Proof. That $(\nabla f)'(x)$ exists almost everywhere has already been shown. In the proof of Alexandrov's theorem we showed that if $F(x) = x + \nabla f(x)$ then the inverse G of F is single valued, Lipschitz, and that almost every point $x \in \mathbf{R}^n$ is of the form $x = G(y)$ at a point where $G'(y)$ is nonsingular and at these points $F'(G(y)) = G'(y)^{-1}$. In Corollary 6.5 we showed that there is a convex function g so that $\nabla g(y) = G(y)$. But G is Lipschitz and therefore its classical derivatives exist almost everywhere and are bounded. Moreover the distributional derivatives of G will equal the classical derivatives. But as $G = \nabla g$ this implies that the classical second derivatives D_iD_jg of g exist almost everywhere and that they are equal to the distributional second derivatives of g. But the (as distributional derivatives commute) we have that the classical derivatives satisfy $D_iD_jg(y) = D_jD_ig(y)$ for almost all y. That is the matrix of $G'(y) = (\nabla g)'(y)$ is symmetric. But then as $F'(G(y)) = G'(y)^{-1}$ the matrix of $F'(G(y))^{-1}$ is symmetric. (Note that the set $A := \{y : G'(y) \text{ is not symmetric}\}$ has measure zero and G is Lipschitz so $G[A]$ will also be of measure zero and so we can ignore the $x = G(y)$ where $G'(y)$ is not symmetric.) This shows $F'(x)$ is symmetric for almost all x. But $F(x) = x + \nabla f(x)$ implies $F'(x) = I + (\nabla f)'(x)$ so that $(\nabla f)'(x)$ is symmetric at exactly the same set of points where $F'(x)$ is symmetric. This shows that $(\nabla f)'(x)$ is symmetric for almost all x.

That $(\nabla f)(x)$ is positive semi-definite follows from a calculation and the fact that f is convex (and thus has a lower supporting linear hyperplane at each point). Details are left to the reader. □

References

1. A. D. Alexandrov, *Almost everywhere existence of the second differential of a convex function and some properties of convex surfaces connected with it*, Leningrad State Univ. Annals [Uchenye Zapiski] Math. Ser. **6** (1939), 3–35.

2. A. S. Besicovitch, *A general form of the covering principle and relative differentiation of additive functions*, Proc. Cambridge Philos. Soc. **41** (1945), 103–110.
3. H. Busemann and W. Feller, *Krümmungsindikatritizen konvexer Flčhen*, Acta Math. **66** (1936), 1–47.
4. M. G. Crandall, H. Ishii, and P.-L. Lions, *Users guide to viscosity solutions of second order partial differential equations*, Bull. Amer. Math. Soc. **27** (1992), no. 1, 1–67.
5. F. Mignot, *Contrôle dans les inéquations variationelles elliptiques*, J. Functional Analysis **22** (1976), no. 2, 130–185.
6. H. Rademacher, *Über partielle und totale Differenzierbarkeit I.*, Math. Ann. **79** (1919), 340–359.
7. R. Schneider, *Convex bodies: The Brunn-Minkowski theory*, Encyclopedia of Mathematics and its Applications, vol. 44, Cambridge University Press, 1993.
8. W. P. Ziemer, *Weakly differentiable functions*, Graduate Texts in Mathematics, vol. 120, Springer-Verlag, New York, 1989.