The Inverse Function Theorem via Newton's Method

Michael Taylor

We aim to prove the following result, known as the inverse function theorem.

Theorem 1. *Let F be a C^k map ($k \geq 1$) from a neighborhood of $p \in \mathbb{R}^n$ to $\mathbb{R}^n$, with $F(p) = q$. Assume the derivative $DF(p)$ is invertible. Then there is a neighborhood $\mathcal{U}$ of p and a neighborhood $\mathcal{V}$ of q such that $F : \mathcal{U} \to \mathcal{V}$ is one-to-one and onto, and $F^{-1} : \mathcal{V} \to \mathcal{U}$ is a C^k map. In particular, $F : \mathcal{U} \to \mathcal{V}$ is a diffeomorphism.*

If we demonstrate the bijection $F : \mathcal{U} \to \mathcal{V}$, with continuous inverse $G = F^{-1}$, and show that

$$DG(q) = DF(p)^{-1}, \tag{1}$$

then we can apply the same argument to any $u \in \mathcal{U}$, with image $v = F(u)$, to get

$$DG(v) = \big(DF(G(v))\big)^{-1}, \quad \forall\, v \in \mathcal{V}.$$

This formula implies G is C^1 if F is C^1, and an inductive argument shows that G is C^k if F is C^k.

To get things started, let us set

$$f(x) = A(F(p + x) - q), \quad A = DF(p)^{-1}, \tag{2}$$

so

$$f(0) = 0, \quad Df(0) = I. \tag{3}$$

We will show that f maps a neighborhood U of 0 one-to-one and onto a neighborhood V of 0, and that

$$Df^{-1}(0) = I. \tag{4}$$

Given this, it readily follows that F maps a neighborhood $\mathcal{U}$ of p one-to-one and onto a neighborhood $\mathcal{V}$ of q, and that $G = F^{-1}$ satisfies (1), so this will prove Theorem 1.

We start with injectivity.

Proposition 2. *Let U be a convex, open neighborhood of 0 in $\mathbb{R}^n$ and $f : U \to \mathbb{R}^n$ a C^1 map satisfying*

$$\|Df(x) - I\| \leq \delta < 1, \quad \forall\, x \in U. \tag{5}$$

Then f is one-to-one.

Proof. Taking $x_1, x_2 \in U$, we have

$$\begin{aligned} f(x_2) - f(x_1) &= \int_0^1 \frac{d}{d\tau} f(\tau x_2 + (1-\tau)x_1)\, d\tau \\ &= \int_0^1 Df(\tau x_2 + (1-\tau)x_1)(x_2 - x_1)\, d\tau \\ &= \Phi(x_1, x_2)(x_2 - x_1), \end{aligned} \tag{6}$$

where

$$\Phi(x_1, x_2) = \int_0^1 Df(\tau x_2 + (1-\tau)x_1)\, d\tau. \tag{7}$$

The hypothesis (5) implies

$$\|\Phi(x_1, x_2) - I\| \le \delta, \tag{8}$$

hence $\Phi(x_1, x_2)$ is invertible, and

$$\|\Phi(x_1, x_2)^{-1}\| \le \frac{1}{1-\delta}. \tag{9}$$

Hence $x_2 - x_1 = \Phi(x_1, x_2)^{-1}(f(x_2) - f(x_1))$, and

$$\|x_1 - x_2\| \le M\|f(x_1) - f(x_2)\|, \tag{10}$$

with

$$M = \frac{1}{1-\delta}. \tag{11}$$

This yields injectivity.

We now take U to be a ball $B_R(0)$ centered at 0, or radius R, small enough that (5) holds. We will impose further restrictions on δ as the argument proceeds. We want to show that if $y \in \mathbb{R}^n$ is close enough to 0, then there exists $x \in U$ such that $f(x) = y$. Furthermore, we want to show that

$$\|x - y\| = o(\|y\|), \tag{12}$$

since this will imply (4).

We produce x as the limit of a sequence (x_k), given by an iterative procedure known as Newton's method. We pick

$$x_0 = 0, \tag{13}$$

and inductively define the sequence by

$$x_{k+1} = x_k + Df(x_k)^{-1}(y - y_k), \quad y_k = f(x_k). \tag{14}$$

Our task is to show that if $\|y\|$ is sufficiently small, x_k is well defined for all k and converges to a limit x satisfying $f(x) = y$. From (3), we have

$$x_1 = y. \tag{15}$$

Since we need to have $x_1 \in U = B_R(0)$, let us impose the requirement

$$\|y\| < \frac{R}{2}. \tag{16}$$

Assuming we have $x_0, \dots, x_{k+1} \in U$ and x_{k+1} given by (14), we want to see whether $f(x_{k+1})$ is closer to y than $f(x_k)$ is. Using (6)–(7), with (x_k, x_{k+1}) in place of (x_1, x_2), we have

$$\begin{aligned} f(x_{k+1}) &= f(x_k) + \Phi(x_k, x_{k+1})(x_{k+1} - x_k) \\ &= f(x_k) + \Phi(x_k, x_{k+1})Df(x_k)^{-1}(y - f(x_k)) \\ &= y + [\Phi(x_k, x_{k+1}) - Df(x_k)]Df(x_k)^{-1}(y - f(x_k)), \end{aligned} \tag{17}$$

the second equality by (14). Hence

$$\|f(x_{k+1}) - y\| \le \varepsilon_k M \|f(x_k) - y\|, \tag{18}$$

where

$$\begin{aligned} \varepsilon_k &= \|\Phi(x_k, x_{k+1}) - Df(x_k)\| \\ &\le \sup_{0 \le \tau \le 1} \|Df(\tau x_{k+1} + (1-\tau)x_k) - Df(x_k)\|. \end{aligned} \tag{18A}$$

If (5) holds, then certainly $\varepsilon_k \le 2\delta$. Note that

$$\delta = \frac{1}{5} \Rightarrow M = \frac{5}{4} \Rightarrow 2\delta M = \frac{1}{2} < 1. \tag{19}$$

Hence we will strengthen (5) to

$$\delta \le \frac{1}{5}. \tag{20}$$

Then (18) implies

$$\|f(x_{k+1}) - y\| \le \frac{1}{2}\|f(x_k) - y\|, \tag{21}$$

hence

$$\begin{aligned} \|f(x_{k+1}) - y\| &\le 2^{-k}\|f(x_1) - y\| \\ &= 2^{-k}\|f(y) - y\|. \end{aligned} \tag{22}$$

To estimate $f(y) - y$, we use (3) and write

$$\begin{aligned} f(y) &= \int_0^1 \frac{d}{d\tau} f(\tau y)\, d\tau \\ &= \int_0^1 Df(\tau y) y\, d\tau \\ &= y + R(y)y, \end{aligned} \tag{23}$$

where

$$R(y) = \int_0^1 Df(\tau y)\, d\tau - Df(0) \tag{24}$$

satisfies

$$\|R(y)\| \le \sup_{0\le\tau\le 1} \|Df(\tau y) - Df(0)\|. \tag{25}$$

We have

$$\|f(y) - y\| \le \|R(y)y\|, \quad \lim_{y\to 0} \|R(y)\| = 0. \tag{26}$$

Hence

$$\|f(x_{k+1}) - y\| \le 2^{-k}\|R(y)y\|. \tag{27}$$

Meanwhile, from (14), we have

$$\begin{aligned} \|x_{k+1} - x_k\| &\le \|Df(x_k)^{-1}\| \cdot \|f(x_k) - y\| \\ &\le 2^{-(k-1)} M \|R(y)y\|, \end{aligned} \tag{28}$$

the latter estimate by (27). Since $x_1 = y$, it follows that

$$\|x_{k+1} - y\| \le K\|R(y)y\|, \quad K = M\sum_{j=0}^{\infty} 2^{-j} = 2M. \tag{29}$$

Thus, if we supplement (16) with the requirement

$$K\|R(y)y\| < \frac{R}{4}, \tag{30}$$

we are guaranteed to have x_ℓ well defined and in U for all $\ell \in \mathbb{Z}^+$. The estimates (28) and (29) imply (x_ℓ) is a Cauchy sequence, with limit $x \in U$, and then (27) implies $f(x) = y$. Furthermore, (29) gives

$$\|x - y\| \le K\|R(y)y\|, \tag{31}$$

so (12) holds, and hence (4) holds. The proof of Theorem 1 is complete.

REMARK. The passage from (18)–(18A) to (21), while giving simpler estimates to work with, loses information. While these simpler estimates are adequate to establish convergence of the sequence (x_k), they fail to show that actually the convergence is dramatically faster than indicated in (27). In fact, for any C^1 function f, (18A) implies

$$\varepsilon_k \to 0 \ \text{ as } \ \|x_{k+1} - x_k\| \to 0, \tag{32}$$

so the estimate

$$\|f(x_{k+1}) - y\| \le \Big(\prod_{j=1}^{k} \varepsilon_j\Big)\|f(y) - y\| \tag{33}$$

is stronger than (22). In case f is C^2, satisfying

$$\|D^2 f(x)\| \le A, \quad \forall\, x \in U, \tag{34}$$

we have

$$\varepsilon_k \le A\|x_{k+1} - x_k\|, \tag{35}$$

which is superior to $\varepsilon_k \le 2\delta$ as soon as

$$\|x_{k+1} - x_k\| \le \frac{2\delta}{A}. \tag{36}$$

Say the first occurrence of (36) is at $k = m$. An upper bound for m can be obtained from (28). Then, for $k \ge m$, we can improve (21) to

$$\begin{aligned} \|f(x_{k+1}) - y\| &\le A\|x_{k+1} - x_k\| \cdot \|f(x_k) - y\| \\ &\le MA\|f(x_k) - y\|^2, \end{aligned} \tag{37}$$

the last inequality of (37) by the first inequality of (28). If $k \ge m$ and m is large enough that

$$MA\|f(x_m) - y\| \le \frac{1}{2}, \tag{38}$$

then the last estimate in (37) implies extremely fast convergence of $f(x_k)$ to $f(y)$, for $m \leq k \nearrow \infty$, and hence, again by the first estimate in (28), correspondingly fast convergence of x_k to x.

REMARK 2. It is clear from the correspondence (2) between F and f that one can implement Newton's method directly on F. If v is sufficiently close to q, one can obtain the solution to $F(u) = v$ as the limit of the sequence (u_k) defined inductively by

$$u_0 = p, \quad u_{k+1} = u_k + DF(u_k)^{-1}(v - v_k), \quad v_k = F(u_k). \tag{39}$$

Let us return to the setting of Proposition 2, and not impose the additional condition (20). Pick positive $S < R$. With Theorem 1 proven, we know that f is a diffeomorphism of the closed ball $\overline{B_S(0)}$ onto a neighborhood $\overline{\Omega}$ of 0,

$$f : \overline{B_S(0)} \longrightarrow \overline{\Omega}. \tag{40}$$

We can use Proposition 2 and some further arguments to show that $\overline{\Omega}$ contains a certain ball $B_\rho(0)$, of radius ρ that we specify shortly. In fact, applying (10) with $x_1 = 0,\ x_2 = x \in \partial B_S(0)$, we have

$$x \in \partial B_S(0) \Longrightarrow \|f(x)\| \geq (1 - \delta)S. \tag{41}$$

With this result in hand, we will prove the following.

Proposition 3. *The image $\overline{\Omega}$ in (40) has the property*

$$\overline{\Omega} \supset B_{(1-\delta)S}(0). \tag{42}$$

That is to say, if $\|y\| < (1 - \delta)S$, then there exists $x \in \overline{B_S(0)}$ such that $f(x) = y$.

Proof. Take $y \in \mathbb{R}^n$ satisfying $\|y\| < (1 - \delta)S$. Set

$$\gamma(t) = ty, \quad 0 \leq t \leq 1. \tag{43}$$

This is a path from 0 to y within the set $B_{(1-\delta)S}(0)$. Assume

$$y \notin \overline{\Omega}, \tag{44}$$

i.e., $\gamma(1) = y$ belongs to the open set $\mathbb{R}^n \setminus \overline{\Omega}$. Meanwhile, $\gamma(0) = 0$ belongs to the open set Ω. It follows that

$$s_0 = \inf\{s \in [0, 1] : \gamma(s) \notin \overline{\Omega}\} \in (0, 1). \tag{45}$$

For $s \in [0, s_0)$, $\gamma(s) \in \overline{\Omega}$, and for each $\varepsilon > 0$ there exists $s \in (s_0, s_0 + \varepsilon)$ such that $\gamma(s) \notin \overline{\Omega}$. These conditions imply

$$\gamma(s_0) \in \partial\Omega. \tag{46}$$

However,

$$f : \partial B_S(0) \longrightarrow \partial\Omega \tag{47}$$

is one-to-one and onto, so (41) implies

$$z \in \partial\Omega \Longrightarrow \|z\| \geq (1-\delta)S. \tag{48}$$

On the other hand,

$$\|\gamma(s_0)\| = |s_0| \cdot \|y\| < (1-\delta)S. \tag{49}$$

The results (46)–(49) are contradictory, so (44) cannot hold. This proves Proposition 3.