Inverse Function Theorem

Ethan Y. Jaffe

1 Motivation

When as an undergraduate I first learned the inverse function theorem, I was using a textbook of Munkres [1]. The proof presented there was not very illuminating to a young analyst, and I never properly learned it.[1] Later, when learning about PDE, an infinite-dimensional version of the inverse function theorem was needed. However, not only was the proof I "knew" not intuitive to me, it also depended essentially on the local compactness of $\mathbf{R}^n$, and so was useless in the infinite-dimensional setting. Later, because of a blog post by Terry Tao [2], I learned that there was always a much easier proof, which extended readily to infinite dimensions. So, partially for myself, I present the details of this argument.

2 Inverse Function Theorem

We will prove the following theorem

Theorem 2.1. *Let U be an open set in $\mathbf{R}^n$, and let $f : U \to \mathbf{R}^n$ be continuously differentiable. Suppose that $x_0 \in U$ and $Df(x_0)$ is invertible. Then there exists a smaller neighbourhood $V \ni x_0$ such that f is a homeomorphism onto its image. Furthermore, V may be taken small enough so that f^{-1} is also continuously differentiable, with its derivative satisfying $D(f^{-1})_y = (Df)^{-1}_{f^{-1}(y)}$. Moreover, if f is of class C^k, ($k \in \mathbf{N} \cup \{\infty\}$), then so is f^{-1}.*

The version of the proof presented here depends on a version of the Banach fixed point theorem with parameter, which we now state.

[1] I would not recommend Munkres' book. The first half, on calculus in Euclidean space, is passable, although the proofs of two of the most important theorems, the invese function theorem and the change of variables theorem, are some of my least favourites in the literature. Although I dislike his proof of the inverse function theorem (hence this note), I especially dislike the proof of the change of variables theorem which completely hides the measure-theoretic intuition behind it. The second half on manifolds is terrible. For whatever misguided reason, Munkres insists on working with manifolds only in $\mathbf{R}^n$, which allows him to conflate the differential structure (i.e. intrinsic) and Riemannian structure (i.e. extrinsic and coming from the embedding into $\mathbf{R}^n$) of a manifold. This conflation is maximally obfuscatory. After learning manifolds this way, it took me a great deal effort to try to untangle the different notions and learn it properly. I would only suggest this book if one wants a belaboured introduction to the Riemann integral in dimension > 1, since this is the only material which I have not seen reproduced better elsewhere (although why exactly one would want to put the effort into the intricacies of the Riemann integral is a different question).

Theorem 2.2 (Banach Fixed Point Theorem)**.** *Let (X, d) be a complete metric space, and $T : X \to X$ be a contraction of factor $r < 1$, i.e. $d(Tx, Ty) \leq rd(x, y)$. Then T has a unique fixed point. Furthermore, if Λ is another metric space, and $T(\lambda)$ is a continuous family of contractions of factor r, continuous in the sense that*

$$\lim_{\lambda \to \lambda_0} \sup_{x \in X} d(T(\lambda)x, T(\lambda_0)x) = 0$$

then the fixed points of $T(\lambda)$ are continuous of λ. Stated otherwise, if $x(\lambda)$ is the unique fixed point of $T(\lambda)$, then the map $\lambda \mapsto x(\lambda)$ is continuous.

Proof. First we show uniqueness. If $Tx = x$ and $Ty = y$, then

$$d(x, y) = d(Tx, Ty) \leq rd(x, y),$$

which is only possible if $d(x, y) = 0$, i.e. $x = y$.

Now for existence. Fix any $x_0 \in X$, and consider $y = \lim_{n\to\infty} T^n(x_0)$. If this exists, then

$$T(y) = T\left(\lim_{n\to\infty} T^n(x_0)\right) = \lim_{n\to\infty} T^{n+1}(x_0) = y,$$

since T is continuous. To prove convergence, notice that the sequence is Cauchy. Indeed, for any n it is easy to see inductively that

$$d(T^n(x_0), T^{n+1}(x_0)) \leq r^n d(x_0, T(x_0)).$$

By the triangle inequality, it follows that for $k \geq 1$

$$d(T^n(x_0), T^{n+k}(x_0)) \leq d(x_0, T(x_0)) \sum_{i=n}^{n+k-1} r^i \leq r^n \frac{d(x_0, T(x_0)}{1 - r}$$

This upper bound is independent of k, so it follows that if $n, m \geq N$, $d(T^n(x_0), T^m(x_0) \leq r^N \frac{d(x_0,T(x_0)}{1-r}$, which shows that the sequence is Cauchy.

Now for the version with parameter. Observe that

$$\begin{aligned} d(x(\lambda), x(\lambda_0)) &= d(T(\lambda)x(\lambda), T(\lambda_0)x(\lambda_0)) \\ &\leq d(T(\lambda)x(\lambda), T(\lambda)x(\lambda_0)) + d(T(\lambda)x(\lambda_0), T(\lambda_0)x(\lambda_0) \\ &\leq rd(x(\lambda), x(\lambda_0)) + d(T(\lambda)(x(\lambda_0)), T(\lambda_0)(x(\lambda_0))). \end{aligned}$$

Rearranging,

$$d(x(\lambda), x(\lambda_0)) \leq (1 - r)^{-1} d(T(\lambda)x(\lambda_0), T(\lambda_0)x(\lambda_0) \to 0$$

as $\lambda \to \lambda_0$ by continuity of the map $\lambda \mapsto T(\lambda)$. □

Now we prove the inverse function theorem.

Proof. Translating and multiplying by a linear map, we may assume that $x_0 = 0, f(x_0) = 0$ and $Df_0 = \mathrm{Id}$. Since f is continuously differentiable, Df_x remains close to Df_0 as matrices if x is close to 0. For $y \in \mathbf{R}^n$, with y close to 0, consider the map $T_y : x \mapsto x - f(x) + y$. Observe that a fixed point x of T_y is precisely an x for which $f(x) = y$.[2]

Let B_R denote the closed ball of radius $R > 0$ centred at 0. B_R is in particular a complete metric space. We will prove that if R is small enough, and y is small enough, T_y maps B_R to itself and is a contraction. We will use $\|x\|$ to denote the usual (ℓ^2) Euclidean norm on points, and for a linear map A, $\|A\|$ to denote the ℓ^2 operator norm.

Let us start by considering the map $F(x) = f(x) - x$. F is continuously differentiable with $DF_0 = 0$. Then for $R > 0$ small enough to that $B_R \subseteq U$, and any two $x, x' \in B_R$,

$$\begin{aligned}\|F(x) - F(x')\| &= \left\| \int_0^1 DF_{(x-x')t+x'} \cdot (x - x')\ dt \right\| \\ &\leq \int_0^1 \|DF_{(x-x')t+x'}\| \|x - x'\|\ dt \\ &\leq \left(\sup_{z \in B_R} \|DF_z\| \right) \|x - x'\|.\end{aligned}$$

Since $DF_0 = 0$ and F is continuously differentiable, for all $0 < \varepsilon < 1$, if R is small enough, $\left(\sup_{z \in B_R} \|DF_z\|\right) \leq \varepsilon$. Fix such an ε.

Suppose $\|y\| \leq R(1 - \varepsilon)$. Then we will show $T_y : B_R \to B_R$ and is a contraction.

Fix $x \in B_R$. Then we compute

$$\begin{aligned}\|T_y(x)\| = \|x - f(x) + y\| &\leq \|F(x)\| + \|y\| \\ &= \|F(x) - F(0)\| + \|y\| \\ &\leq \varepsilon \|x\| + R(1 - \varepsilon) \leq R.\end{aligned}$$

[2]Let us motivate the choice of this map. Let us use the notation $F(x) = f(x) - x$ of the sequel. Since $Df_0 = \mathrm{Id}$, $F(x) \in o(1)$ can be thought of as a perturbation of the constant map 0, and hence $f = \mathrm{id} - F$ is a peturbation of the identity. We are seeking to solve $f(x) = y$, i.e. $(\mathrm{id} - F)(x) = y$. One way to motivate the choice of the map T_y is to rearrange this equality into $\mathrm{id}(x) = F(x) + y = T_y(x)$, i.e. finding a fixed point for T_y. A more brute force approach, however, is to try to build a sequence of approximate solutions x_n to this equation, starting with $x_1 = y$, and iteratively improving the error. Explicitly, we may think of $F(x_n)+y = x_n+R_n$, where R_n is some error. To improve the error, we try to perturb by adding some z of size roughly R_n to x_n, and trying to solve $F(x_n+z)+y = x_n+z$. For such z, $F(x_n+z) = F(x_n)+DF_{x_n}z+o(|z|)$ by definition of the derivative, and so we are trying to solve $F(x_n) + y + DF_{x_n}z + o(|z|) = x_n + z$, i.e. $R_n + DF_{x_n}z + o(|z|) = z$. Since x_n should be thought of as close to 0, DF_{x_n} is small, and thus $DF_{x_n}z + o(|z|)$ should be thought of as neglible compared to R_n if z is about size R_n. Thus $z = R_n$ is of size R_n and solves the previous equation modulo an error of $R_{n+1} = DF_{x_n}R_n + o(|R_n|)$, which is of smaller order than R_n. Thus $x_{n+1} := x_n + z$ solves $F(x_{n+1}) + y = x_{n+1} + R_{n+1}$, and R_{n+1} is an improved error compared to R_n.

While this idea can be turned into a formal proof, one should just notice that $x_{n+1} = x_n + z = x_n + R_n = F(x_n) + y = T_y(x_n)$ is just a fixed-point iteration, and so the formalization of this proof may be abstracted away to an invokation of the fixed point theorem, anyway. The reader should also notice the similary of this approach to inverting the linear operator $\mathrm{Id} - F$ on a Banach space, if $\|F\| < 1$, via the Neumann series $\sum_{k=0}^{\infty} F^k$; in this case, the choice of x_n above are precisely the partial sums of the Neumann series, truncated at $k = 2^{n-1}$, applied to y.

Thus $T_y : B_R \to B_R$.

Now for the contraction. Fix $x, x' \in B_R$. Then we compute

$$\|T_y(x) - T_y(x')\| \leq \|F(x) - F(x')\| \leq \varepsilon \|x - x'\|.$$

By the fixed point theorem, T_y has a unique fixed point $x \in B_R$, i.e. if $\|y\|$ is small enough, there exists a unique solution x to $f(x) = y$ with $x \in B_R$. In other words, we have established the existence of $f^{-1} : B_{R(1-\varepsilon)} \to B_R$.

We still need to prove that f is a homeomorphism. In finite dimensions, we can appeal to the fact that a continuous bijection between compact subsets of $\mathbf{R}^n$ is a homeomorphism. The main purpose of this note is to show that we do not need the assumption of finite dimensions, so we will use the version of the fixed point theorem with parameter. So, we just need to prove that f^{-1} is continuous, i.e. the fixed points of T_y are continuous in y. By the fixed point theorem, we just need to show that the map $y \to T_y$ is continuous, since they all have the same contractive factor ε. We easily compute for $y, y_0 \in B_{R(1-\varepsilon)}$.

$$\sup_{x \in B_R} \|T_y x - T_{y_0} x\| = \|y - y_0\|,$$

which certainly tends to 0 as $y \to y_0$. Thus f^{-1} is continuous. If $0 \in V \subseteq B_R$ is open, then restricting f to U, it follows that f is a homeomorphism onto its image, which we will cal W. This completes the first part of the theorem

Now we need to show that f^{-1} is continuously differentiable. Shrinking V if necessary, we may assume that Df_x is nonsingular on V. Now we show that $f^{-1} : W \to V$ (which we know to be a homeomorphism) is differentiable on W, with derivative $(Df)^{-1}_{f^{-1}(y)}$. Since Df is non-singular and f^{-1} is continuous, this automatically shows that $(Df)^{-1}_{f^{-1}(y)}$ is continuous, and hence f^{-1} is continuously differentiable. Fix $y_0 \in W$, and write $x_0 = f^{-1}(y_0)$, and for any $y \in W$ write $x = f^{-1}(y)$. Then since f is a homeomorphism

$$\begin{aligned}
&\lim_{y \to y_0} \frac{f^{-1}(y) - f^{-1}(y_0) - (Df)^{-1}_{f^{-1}(y_0)}(y - y_0)}{\|y - y_0\|} \\
&= \lim_{x \to x_0} \frac{x - x_0 - (Df)^{-1}_{x_0}(f(x) - f(x_0))}{\|f(x) - f(x_0)\|} \\
&= \lim_{x \to x_0} -Df^{-1}_{x_0} \left(\frac{f(x) - f(x_0) - Df_{x_0}(x - x_0)}{\|x - x_0\|} \right) \frac{\|x - x_0\|}{\|f(x) - f(x_0)\|}
\end{aligned}$$

Since $Df^{-1}_{x_0}$ is a linear map, it is continuous, and so the first factor converges to 0 by definition of differentiability. The second factor is bounded above as $x \to x_0$. Indeed,

$$\begin{aligned}
\liminf_{x \to x_0} \frac{\|f(x) - f(x_0)\|}{\|x - x_0\|} &\geq \liminf_{x \to x_0} \left| \frac{\|Df_{x_0}(x - x_0)\|}{\|x - x_0\|} - \frac{\|f(x) - f(x_0) - Df_{x_0}(x - x_0)\|}{\|x - x_0\|} \right| \\
&= \liminf_{x \to x_0} \frac{\|Df_{x_0}(x - x_0)\|}{\|x - x_0\|} \geq c > 0,
\end{aligned}$$

since Df_{x_0} invertible means that there is some $c > 0$ for which $\|Df_{x_0}(x - x_0)\| \geq c\|x - x_0\|$. Putting these two things together means that

$$\lim_{y \to y_0} \frac{f^{-1}(y) - f^{-1}(y_0) - (Df)^{-1}_{f^{-1}(y_0)}(y - y_0)}{\|y - y_0\|} = 0,$$

i.e. f^{-1} is differentiable at y_0 with the desired derivative.

Lastly, we show that if f is C^k on V, then f^{-1} is C^k on W, *without* the need to shrink V. Because we do not shrink V, if we can show this is true for $k < \infty$, we automatically show it's true for $k = \infty$. First, observe that $\mathrm{GL}(n, \mathbf{R})$ is an open subset of $M_n(\mathbf{R}) \cong \mathbf{R}^{n^2}$, and that the inversion map $I : \mathrm{GL}(n, \mathbf{R}) \to \mathrm{GL}(n, \mathbf{R})$ is of class C^∞ (since it is just a rational function of the entries). If f is of class C^k, then the map $Df : V \to \mathrm{GL}(n, \mathbf{R})$ is of class C^{k-1}. Now, from the above, $D(f^{-1}) : W \to \mathrm{GL}(n, \mathbf{R})$ is just

$$D(f^{-1}) = I \circ Df \circ f^{-1},$$

i.e. is the composition of three maps, the first of which is C^∞, and the second of which is C^{k-1}. This argument shows that that if f^{-1} is of class C^r for $r < k$, then $D(f^{-1})$ is of class C^r, too, so that f^{-1} is of class C^{r+1}. Starting with the case $r = 1$, which we know to be true, we obtain iteratively that f^{-1} is of class C^k, too. □

Remark 2.3. As mentioned in the motivation section, this proof easily extends to infinite dimensions, with the derivative replaced by the Fréchet derivative. Indeed, the only thing which needs changing is $\mathbf{R}^n$ to whichever Banach space X is in question, and changing the norms to the norms in the Banach spaces. The last part about f^{-1} inheriting the regularity of f does not quite carry through, but this is alright for the purpose of this note, since anyway it is not clear what it means for a map to be k times continuously Fréchet differentiable.

References

[1] James R Munkres, *Analysis on Manifolds*

[2] Terence Tao, *The inverse function theorem for everywhere differentiable maps*. https://terrytao.wordpress.com/2011/09/12/the-inverse-function-theorem-for-everywhere-differentiable-maps/