Appendix B

# The Magic of Iteration

The subject of this appendix is one of our favorites in all of mathematics, and it's not hard to explain why. As you will see, the basic theorem, the Banach Contraction Principle, has a simple and elegant statement and a proof to match. And yet, at the same time, it is extremely powerful, having as easy consequences two of the most important foundations of advanced analysis, the Implicit Function Theorem and the Local Existence and Uniqueness Theorem for systems of ODE.

But there is another aspect that we find very appealing, and that is that the basic technique that goes into the contraction principle, namely iteration of a mapping, leads to remarkably simple and effective algorithms for solving equations. Indeed what the Banach Contraction Principle teaches us is that if we have a good algorithm for evaluating a function $f(x)$, then we can often turn it into an algorithm for inverting $f$, i.e., for solving $f(x) = y$!

## B.1. The Banach Contraction Principle

In what follows we will assume that $X$ is a metric space and that $f : X \to X$ is a continuous mapping of $X$ to itself. Since $f$ maps $X$ to itself, we can compose $f$ with itself any number of times, so we can define $f^0(x) = x$, $f^1(x) = f(x)$, $f^2(x) = f(f(x))$, and inductively $f^{n+1}(x) = f(f^n(x))$. The sequence $f^n(x)$ is called the sequence of iterates of $x$ under $f$, or the orbit of $x$ under $f$. By associativity of composition, $f^n(f^m(x)) = f^{n+m}(x)$, and by Exercise A-1 of Appendix A, if $K$ is a Lipschitz constant for $f$, then $K^n$ is a Lipschitz

constant for $f^n$. We shall use both of these facts below without further mention.

A point $x$ of $X$ is called a fixed point of $f$ if $f(x) = x$. Notice that finding a fixed point amounts to solving a special kind of equation. What may not be obvious is that solving many other types of equations can often be reduced to solving a fixed-point equation. We will give other examples later, but here is a typical reduction. Assume that $V$ is a vector space and that we want to solve the equation $g(x) = y$ for some (usually nonlinear) map $g : V \to V$. Define a new map $f : V \to V$ by $f(x) = x - g(x) + y$. Then clearly $x$ is a fixed point of $f$ if and only if it solves $g(x) = y$. This is in fact the trick used to reduce the Inverse Function Theorem to the Banach Contraction Principle.

The Banach Contraction Principle is a very general technique for finding fixed points. First notice the following: if $x$ is a point of $X$ such that the sequence $f^n(x)$ of iterates of $x$ converges to some point $p$, then $p$ is a fixed point of $f$. In fact, by the continuity of $f$, $f(p) = f(\lim_{n\to\infty} f^n(x)) = \lim_{n\to\infty} f(f^n(x)) = \lim_{n\to\infty} f^{n+1}(x) = p$. We will see that if $f$ is a contraction, then for any point $x$ of $X$ the sequence of iterates of $x$ is in any case a Cauchy sequence, so if $X$ is complete, then it converges to a fixed point $p$ of $f$. In fact, we will see that a contraction can have at most one fixed point $p$, and so to locate this $p$ when $X$ is complete, we can start at any point $x$ and "follow the iterates of $x$ to their limit". This in essence is the Banach Contraction Principle. Here are the details.

**B.1.1. Fundamental Contraction Inequality.** *If $f : X \to X$ is a contraction mapping and if $K < 1$ is a Lipschitz constant for $f$, then for all $x_1$ and $x_2$ in $X$,*

$$\rho(x_1, x_2) \le \frac{1}{1-K}(\rho(x_1, f(x_1)) + \rho(x_2, f(x_2))).$$

**Proof.** The triangle inequality,

$$\rho(x_1, x_2) \le \rho(x_1, f(x_1)) + \rho(f(x_1), f(x_2)) + \rho(f(x_2), x_2),$$

together with $\rho(f(x_1), f(x_2)) \leq K\rho(x_1, x_2)$ gives

$$\rho(x_1, x_2) - K\rho(x_1, x_2) \leq \rho(x_1, f(x_1)) + \rho(f(x_2), x_2).$$

Since $1 - K > 0$, the desired inequality follows. ∎

This is a very strange inequality: it says that we can estimate how far apart any two points $x_1$ and $x_2$ are just from knowing how far $x_1$ is from its image $f(x_1)$ and how far $x_2$ is from its image $f(x_2)$. As a first application we have

**B.1.2. Corollary.** *A contraction can have at most one fixed point.*

**Proof.** If $x_1$ and $x_2$ are both fixed points, then $\rho(x_1, f(x_1))$ and $\rho(x_2, f(x_2))$ are zero, so by the Fundamental Inequality $\rho(x_1, x_2)$ is also zero. ∎

**B.1.3. Proposition.** *If $f : X \to X$ is a contraction mapping, then, for any $x$ in $X$, the sequence $f^n(x)$ of iterates of $x$ under $f$ is a Cauchy sequence.*

**Proof.** Taking $x_1 = f^n(x)$ and $x_2 = f^m(x)$ in the Fundamental Inequality gives

$$\rho(f^n(x), f^m(x)) \leq \frac{1}{1-K}(\rho(f^n(x), f^n(f(x))) + \rho(f^m(x), f^m(f(x)))).$$

Since $K^n$ is a Lipschitz constant for $f^n$,

$$\rho(f^n(x), f^m(x)) \leq \frac{K^n + K^m}{1-K}\rho(x, f(x)),$$

and since $0 \leq K < 1$, $K^n \to 0$, so $\rho(f^n(x), f^m(x)) \to 0$ as $n$ and $m$ tend to infinity. ∎

**B.1.4. Banach Contraction Principle.** *If $X$ is a complete metric space and $f : X \to X$ is a contraction mapping, then $f$ has a unique fixed point $p$, and for any $x$ in $X$ the sequence $f^n(x)$ converges to $p$.*

**Proof.** The proof is immediate from the above. ∎

▷ **Exercise B–1.** Use the mean value theorem of differential calculus to show that if $X = [a, b]$ is a closed interval and $f : X \to R$ is a continuously differentiable real-valued function on $X$, then the maximum value of $|f'|$ is the smallest possible Lipschitz constant for $f$. In particular $\sin(1)$ (which is less than 1) is a Lipschitz constant for the cosine function on the interval $X = [-1, 1]$. Note that for any $x$ in $R$ the iterates of $x$ under cosine are all in $X$. Deduce that no matter where you start, the successive iterates of cosine will always converge to the same limit. Put your calculator in radian mode, enter a random real number, and keep hitting the cos button. What do the iterates converge to?

As the above exercise suggests, if we can reinterpret the solution of an equation as the fixed point of a contraction mapping, then it is an easy matter to write an algorithm to find it. Well, almost—something important is still missing, namely, when should we stop iterating and take the current value as the "answer"? One possibility is to just keep iterating until the distance between two successive iterates is smaller than some predetermined "tolerance" (perhaps the machine precision). But this seems a little unsatisfactory, and there is actually a much neater "stopping rule".

Suppose we are willing to accept an "error" of $\epsilon$ in our solution; i.e., instead of the actual fixed point $p$ of $f$ we will be happy with any point $p'$ of $X$ satisfying $\rho(p, p') < \epsilon$. Suppose also that we start our iteration at some point $x$ in $X$. It turns out that it is easy to specify an integer $N$ so that $p' = f^N(x)$ will be a satisfactory answer. The key, not surprisingly, lies in the Fundamental Inequality, which we apply now with $x_1 = f^N(x)$ and $x_2 = p$. It tells us that $\rho(f^N(x), p) \leq \frac{1}{1-K}\rho(f^N(x), f^N(f(x))) \leq \frac{K^N}{1-K}\rho(x, f(x))$. Since we want $\rho(f^N(x), p) \leq \epsilon$, we just have to pick $N$ so large that $\frac{K^N}{1-K}\rho(x, f(x)) < \epsilon$. Now the quantity $d = \rho(x, f(x))$ is something that we can compute after the first iteration and we can then compute how large $N$ has to be by taking the log of the above inequality and solving for $N$ (remembering that $\log(K)$ is negative). We can express our result as

**B.1.5. Stopping Rule.** *If* $d = \rho(x, f(x))$ *and*

$$N > \frac{\log(\epsilon) + \log(1-K) - \log(d)}{\log(K)},$$

*then* $\rho(f^N(x), p) < \epsilon$.

From a practical programming point of view, this allows us to express our iterative algorithm with a "for loop" rather than a "while loop", but this inequality has another interesting interpretation. Suppose we take $\epsilon = 10^{-m}$ in our stopping rule inequality. What we see is that the growth of $N$ with $m$ is a constant plus $m/|\log(K)|$, or in other words, to get one more decimal digit of precision we have to do (roughly) $1/|\log(K)|$ more iteration steps. Stated a little differently, if we need $N$ iterative steps to get $m$ decimal digits of precision, then we need another $N$ to double the precision to $2m$ digits.

We say a numerical algorithm has linear convergence if it exhibits this kind of error behavior, and if you did the exercise above for locating the fixed point of the cosine function, you would have noticed it was indeed linear. Linear convergence is usually considered somewhat unsatisfactory. A *much* better kind of convergence is quadratic, which means that each iteration should (roughly) double the number of correct decimal digits. Notice that the actual linear rate of convergence predicted by the above stopping rule is $1/|\log(K)|$. So one obvious trick to get better convergence is to see to it that the best Lipschitz constant for our iterating function $f$ in a neighborhood of the fixed point $p$ actually approaches zero as the diameter of the neighborhood goes to zero. If this happens at a fast enough rate, we may even achieve quadratic convergence, and that is what actually occurs in "Newton's Method", which we study next.

▷ **Exercise B–2.** Newton's Method for finding $\sqrt{2}$ gives the iteration $x_{n+1} = x_n/2 + 1/x_n$. Start with $x_0 = 1$, and carry out a few steps to see the impressive effects of quadratic convergence.

**B.1.6. Remark.** Suppose $V$ and $W$ are orthogonal vector spaces, $U$ is a convex open set in $V$, and $f : U \to W$ is a continuously differentiable map. Let's try to generalize the exercise above to find a

Lipschitz constant for $f$. If $p$ is in $U$, then recall that $Df_p$, the differential of $f$ at $p$, is a linear map of $V$ to $W$ defined by $Df_p(v) = (d/dt)_{t=0} f(p+tv)$, and it then follows that if $\sigma(t)$ is any smooth path in $U$, then $d/dt f(\sigma(t)) = Df_{\sigma(t)}(\sigma'(t))$. If $p$ and $q$ are any two points of $U$ and if $\sigma(t) = p+t(q-p)$ is the line joining them, then integrating the latter derivative from 0 to 1 gives the so-called "finite difference formula": $f(q)-f(p) = \int_0^1 Df_{\sigma(t)}(q-p)\,dt$. Now recall that if $T$ is any linear map of $V$ to $W$, then its norm $\|T\|$ is the smallest nonnegative real number $r$ so that $\|Tv\| \leq r\,\|v\|$ for all $v$ in $V$. Since $\left\|\int_a^b g(t)\,dt\right\| \leq \int_a^b \|g(t)\|\,dt$, $\|f(q) - f(p)\| \leq (\int_0^1 \left\|Df_{\sigma(t)}\right\|\,dt)\,\|(q-p)\|$, and it follows that the supremum of $\|Df_p\|$ for $p$ in $U$ is a Lipschitz constant for $f$. (In fact, it is the smallest one.)

## B.2. Newton's Method

The algorithm called "Newton's Method" has proved to be an extremely valuable tool with countless interesting generalizations, but the first time one sees the basic idea explained, it seems so utterly obvious that it is hard to be very impressed.

Suppose $g : R \to R$ is a continuously differentiable real-valued function of a real variable and $x_0$ is an "approximate root" of $g$, in the sense that there is an actual root $p$ of $g$ close to $x_0$. Newton's Method says that to get an even better approximation $x_1$ to $p$, we should take the point where the tangent line to the graph of $g$ at $x_0$ meets the $x$-axis, namely $x_1 = x_0 - g(x_0)/g'(x_0)$. Recursively, we can then define $x_{n+1} = x_n - g(x_n)/g'(x_n)$ and get the root $p$ as the limit of the resulting sequence $\{x_n\}$.

Typically one illustrates this with some function like $g(x) = x^2-2$ and $x_0 = 1$ (see the exercise above). But the simple picture in this case hides vast difficulties that could arise in other situations. The $g'(x_0)$ in the denominator is a tip-off that things are not going to be simple. Even if $g'(x_0)$ is different from zero, $g'$ could still vanish several times (even infinitely often) between $x_0$ and $p$. In fact, determining the exact conditions under which Newton's Method "works" is a subject in itself, and generalizations of this problem constitute an interesting and lively branch of discrete dynamical systems theory.

We will not go into any of these interesting but complicated questions, but rather content ourselves with showing that under certain simple circumstances we can derive the correctness of Newton's Method from the Banach Contraction Principle.

It is obvious that the right function $f$ to use in order to make the Contraction Principle iteration reduce to Newton's Method is $f(x) = x - g(x)/g'(x)$ and that a fixed point of this $f$ is indeed a root of $g$. On the other hand it is clear that this cannot work if $g'(p) = 0$, so we will assume that $p$ is a "simple root" of $g$, i.e., that $g'(p) \neq 0$. Given $\delta > 0$, let $N_\delta(p) = \{x \in R \mid |x - p| \leq \delta\}$. We will show that if $g$ is $C^2$ and $\delta$ is sufficiently small, then $f$ maps $X = N_\delta(p)$ into itself and is a contraction on $X$. Of course we choose $\delta$ so small that $g'$ does not vanish on $X$, so $f$ is well-defined on $X$. It will suffice to show that $f$ has a Lipschitz constant $K < 1$ on $X$, for then if $x \in X$, then

$$|f(x) - p| = |f(x) - f(p)| \leq K|x - p| < \delta,$$

so $f(x)$ is also in $X$.

But, by one of the exercises, to prove that $K$ is a Lipschitz bound for $f$ in $X$, we only have to show that $|f'(x)| \leq K$ in $X$. Now an easy calculation shows that $f'(x) = g(x)g''(x)/g'(x)^2$. Since $g(p) = 0$, it follows that $f'(p) = 0$ so, by the evident continuity of $f'$, given any $K > 0$, $|f'(x)| \leq K$ in $X$ if $\delta$ is sufficiently small.

The fact that the best Lipschitz bound goes to zero as we approach the fixed point is a clue that we should have better than linear convergence with Newton's Method, but quadratic convergence is not quite a consequence. Here is the proof of that.

Let $C$ denote the maximum of $|f''(x)|$ for $x$ in $X$. Since $f(p) = p$ and $f'(p) = 0$, Taylor's Theorem with Remainder gives $|f(x) - p| \leq C|x - p|^2$. This just says that the error after $n + 1$ iterations is essentially the square of the error after $n$ iterations.

Generalizing Newton's Method to find zeros of a $C^2$ map $G : \mathbf{R}^n \to \mathbf{R}^n$ is relatively straightforward. Let $x_0 \in \mathbf{R}^n$ be an approximate zero of $G$, again in the sense that there is a $p$ close to $x$ with $G(p) = 0$. Let's assume now that $DG_p$, the differential of $G$ at $p$, is nonsingular and hence that $DG_x$ is nonsingular for $x$ near

$p$. The natural analogue of Newton's Method is to define $x_{n+1} = x_n - DG_{x_n}^{-1}(G(x_n))$, or in other words to consider the sequence of iterates of the map $F : N_\delta(p) \to \mathbf{R}^n$ given by $F(x) = x - DG_x^{-1}(G(x))$. Again it is clear that a fixed point of $F$ is a zero of $G$, and an argument analogous to the one-dimensional case shows that for $\delta$ sufficiently small $F : N_\delta(p) \to N_\delta(p)$ is a contraction.

## B.3. The Inverse Function Theorem

Let $V$ and $W$ be orthogonal vector spaces and $g : V \to W$ a $C^k$ map, $k > 0$. Suppose that for some $v_0$ in $V$ the differential $Dg_{v_0}$ of $g$ at $v_0$ is a linear isomorphism of $V$ with $W$. Then the Inverse Function Theorem says that $g$ maps a neighborhood of $v_0$ in $V$ one-to-one onto a neighborhood $U$ of $g(v_0)$ in $W$ and that the inverse map from $U$ into $V$ is also $C^k$.

It is easy to reduce to the case that $v_0$ and $g(v_0)$ are the respective origins of $V$ and $W$, by replacing $g$ by $v \mapsto g(v + v_0) - g(v_0)$. We can then further reduce to the case that $W = V$ and $Dg_0$ is the identity mapping $I$ of $V$ by replacing this new $g$ by $(Dg_0)^{-1} \circ g$.

Given $y$ in $V$, define $f = f_y : V \to V$ by $f(v) = v - g(v) + y$. Note that a solution of the equation $g(x) = y$ is the same thing as a fixed point of $f$. We will show that if $\delta$ is sufficiently small, then $f$ restricted to

$$X = N_\delta = \{v \in V |\ \|v\| \leq \delta\}$$

is a contraction mapping of $N_\delta$ to itself provided $\|y\| < \delta/2$. By the Banach Contraction Principle it then follows that $g$ maps $N_\delta$ one-to-one into $V$ and that the image covers the neighborhood of the origin $U = \{v \in V |\ \|v\| < \delta/2\}$. This proves the Inverse Function Theorem except for the fact that the inverse mapping of $U$ into $V$ is $C^k$, which we will not prove.

The first thing to notice is that since $Dg_0 = I$, $Df_0 = 0$ and hence, by the continuity of $Df$, $\|Df_v\| < 1/2$ for $v$ in $N_\delta$ provided $\delta$ is sufficiently small. Since $N_\delta$ is convex, by a remark above, this proves that $1/2$ is a Lipschitz bound for $f$ in $N_\delta$ and in particular that $f$ restricted to $N_\delta$ is a contraction. Thus it only remains to show that $f$ maps $N_\delta$ into itself provided $\|y\| < \delta/2$. That is, we must

show that if $\|x\| \le \delta$, then also $\|f(x)\| \le \delta$. But since $f(0) = y$,

$$\begin{aligned}\|f(x)\| &\le \|f(x) - f(0)\| + \|f(0)\| \\ &\le \frac{1}{2}\|x\| + \|y\| \\ &\le \delta/2 + \delta/2 \le \delta. \qquad \blacksquare\end{aligned}$$

▷ **Exercise B–3.** The first (and main) step in proving that the inverse function $h : U \to V$ is $C^k$ is to prove that $h$ is Lipschitz. That is, we want to find a $K > 0$ so that given $y_1$ and $y_2$ with $\|y_i\| < \delta/2$ and $x_1$ and $x_2$ with $\|x_i\| < \delta$, if $h(y_i) = x_i$, then $\|x_1 - x_2\| \le K\|y_1 - y_2\|$. Prove this with $K = 2$, using the facts that $h(y_i) = x_i$ is equivalent to $f_{y_i}(x_i) = x_i$ and $1/2$ is a Lipschitz constant for $h = I - g$.

## B.4. The Existence and Uniqueness Theorem for ODE

Let $V : \mathbf{R}^n \times \mathbf{R} \to \mathbf{R}^n$ be a $C^1$ time-dependent vector field on $\mathbf{R}^n$. In the following $I = [a, b]$ will be a closed interval that contains $t_0$ and we will denote by $C(I, \mathbf{R}^n)$ the vector space of continuous maps $\sigma : I \to \mathbf{R}^n$ and define a distance function on $C(I, \mathbf{R}^n)$ by

$$\rho(\sigma_1, \sigma_2) = \max_{t \in I} \|\sigma_1(t) - \sigma_2(t)\|.$$

It is not hard to show that $C(I, \mathbf{R}^n)$ is a complete metric space. In fact, this just amounts to the theorem that a uniform limit of continuous functions is continuous.

Define for each $v_0$ in $\mathbf{R}^n$ a map $F = F^{V,v_0} : C(I, \mathbf{R}^n) \to C(I, \mathbf{R}^n)$ by $F(\sigma)(t) := v_0 + \int_{t_0}^{t} V(\sigma(s), s)\, ds$. The Fundamental Theorem of Calculus gives $\frac{d}{dt}(F(\sigma)(t)) = V(\sigma(t), t)$, and clearly $F(\sigma)(t_0) = v_0$. It follows that if $\sigma$ is a fixed point of $F$, then it is a solution of the ODE $\sigma'(t) = V(\sigma(t), t)$ with initial condition $v_0$, and the converse is equally obvious. Thus it is natural to try to find a solution of this differential equation with initial condition $v_0$ by starting with the constant path $\sigma_0(t) \equiv v_0$ and applying successive approximations using the function $F$. We will now see that this idea works and leads to the following result, called the Local Existence and Uniqueness Theorem for $C^1$ ODE.

**B.4.1. Theorem.** *Let $V : \mathbf{R}^n \times \mathbf{R} \to \mathbf{R}^n$ be a $C^1$ time-dependent vector field on $\mathbf{R}^n$, $p \in \mathbf{R}^n$, and $t_0 \in \mathbf{R}$. There are positive constants $\epsilon$ and $\delta$ depending on $V$, $p$, and $t_0$ such that if $I = [t_0 - \delta, t_0 + \delta]$, then for each $v_0 \in V$ with $\|v_0 - p\| < \epsilon$ the differential equation $\sigma'(t) = V(\sigma(t), t)$ has a unique solution $\sigma : I \to \mathbf{R}^n$ with $\sigma(t_0) = v_0$.*

**Proof.** If $\epsilon > 0$, then, using the technique explained earlier, we can find a Lipschitz constant $M$ for $V$ restricted to the set of $(x, t) \in \mathbf{R}^n \times \mathbf{R}$ such that $\|x - p\| \le 2\epsilon$ and $|t - t_0| \le \epsilon$. Let $B$ be the maximum value of $F(x, t)$ on this same set, and choose $\delta > 0$ so that $K = M\delta < 1$ and $B\delta < \epsilon$, and define $X$ to be the set of $\sigma$ in $C(I, V)$ such that $\|\sigma(t) - p\| \le 2\epsilon$ for all $|t| \le \delta$. It is easy to see that $X$ is closed in $C(I, V)$ and hence a complete metric space. The theorem will follow from the Banach Contraction Principle if we can show that for $\|v_0\| < \epsilon$, $F^{V,v_0}$ maps $X$ into itself and has $K$ as a Lipschitz bound.

If $\sigma \in X$, then $\|F(\sigma)(t) - p\| \le \|v_0 - p\| + \int_0^t \|V(\sigma(s), s)\|\, ds \le \epsilon + \delta B \le 2\epsilon$, so $F$ maps $X$ to itself. And if $\sigma_1, \sigma_2 \in X$, then $\|V(\sigma_1(t), t) - V(\sigma_2(t), t)\| \le M \|\sigma_1(t) - \sigma_2(t)\|$, so

$$\begin{aligned}
\|F(\sigma_1)(t) - F(\sigma_2)(t)\| &\le \int_0^t \|V(\sigma_1(s), s) - V(\sigma_2(s), s)\|\, ds \\
&\le \int_0^t M \|\sigma_1(s) - \sigma_2(s)\|\, ds \\
&\le \int_0^t M\rho(\sigma_1, \sigma_2)\, ds \\
&\le \delta M \rho(\sigma_1, \sigma_2) \le K\rho(\sigma_1, \sigma_2)
\end{aligned}$$

and it follows that $\rho(F(\sigma_1), F(\sigma_2)) \le K\rho(\sigma_1, \sigma_2)$. ∎