The Magic of Iteration

Richard S. Palais
Department of Mathematics
Brandeis University
Waltham, MA 02254-9110

Section 1
Fixed Points, Iteration, and Equation Solving

The subject of these notes is one of my favorites in all mathematics, and it's not hard to say 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 I 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. As my son, Bob Palais, pointed out to me once, what the Banach Contraction Principle teaches us is that if you have a good algorithm for evaluating a function $f(x)$ then you have a good algorithm for solving $f(x) = y$!

1.1 The Banach Contraction Principle.

The natural setting for this subject is a function $f : X \to X$ on a complete metric space X. If you don't know what a complete metric space is don't worry—you can think of X as being R^n with its usual distance function $\rho(x, y) = \|x - y\|$, and you'll get most of the essentials.

But in case you want to know, the next few paragraphs are a Short Course on Metric Spaces that will tell you all you need to know for current purposes.

A metric space is just a set with a *distance* $\rho(x_1, x_2)$ defined between any pair of its points x_1 and x_2. This distance should be a non-negative real number that is zero if and only if $x_1 = x_2$, and it should be symmetric in x_1 and x_2. Aside from these obvious properties of anything we would consider calling a distance function, the only other property we demand of the function ρ (which is also called the *metric* of X) is that the "triangle inequality" hold for any three points x_1, x_2, and x_3 of X. This just means that $\rho(x_1, x_3) \leq \rho(x_1, x_2) + \rho(x_2, x_3)$.

If $\{x_n\}$ is a sequence of points in X, then of course we say this sequence converges to a point x in X if $\lim_{n\to\infty} \rho(x_n, x) = 0$. It is easy to check that the limit x is unique if it exists, and we write $\lim_{n\to\infty} x_n = x$. The sequence $\{x_n\}$ is called a *Cauchy sequence* if $\rho(x_n, x_m)$ converges to zero as both m and n tend to infinity. It is easy to check that a convergent sequence is Cauchy, but the reverse need not be true, and if every Cauchy sequence in X does in fact converge then we call X a *complete* metric space.

If X and Y are metric spaces, and $f : X \to Y$ is a function, then we call f continuous if and only if $f(x_n)$ converges to $f(x)$ whenever x_n converges to x. An equivalent definition is that given any positive ϵ and an x in X, there exists a $\delta > 0$ so that if $\rho_X(x, x') < \delta$ then $\rho_Y(f(x), f(x')) < \epsilon$, and if we can choose this δ independent of x then we call f *uniformly continuous*. A positive constant K is called a Lipschitz constant for f if for all x_1, x_2 in X, $\rho_Y(f(x_1), f(x_2)) < K\rho_X(x_1, x_2)$, and we call f a *contraction* if it has a Lipschitz constant K satisfying $K < 1$. Note that if f has a Lipschitz constant (in particular, if it is a contraction) then f is automatically uniformly continuous (take $\delta = \epsilon/K$).

▷ Exercise 1. Show that if K is a Lipschitz constant for $f : X \to Y$ and L is a Lipschitz constant for $g : Y \to Z$, then KL is a Lipschitz constant for $g \circ f : X \to Z$.

That finishes the Short Course on Metric Spaces.

From here on 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 of points $f^n(x)$ is called the iterates of x under f, or sometimes the orbit of x under f.

We note that, by associativity of composition, $f^n(f^m(x)) = f^{n+m}(x)$. Also, by the above Exercise, 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'll 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 non-linear) 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.

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,*

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

PROOF. The triangle inequality, $\rho(x_1, x_2) \leq \rho(x_1, f(x_1)) + \rho(f(x_1), f(x_2)) + \rho(f(x_2), x_2)$ plus the relation $\rho(f(x_1), f(x_2)) \leq K\rho(x_1, x_2)$ give $\rho(x_1, x_2) - K\rho(x_1, x_2) \leq \rho(x_1, f(x_1)) + \rho(f(x_2), x_2)$. 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:

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. ∎

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 $\rho(f^n(x), f^m(x)) \leq \frac{1}{1-K}(\rho(f^n(x), f^n(f(x))) + \rho(f^m(x), f^m(f(x)))$. Recalling that K^n is a Lipschitz constant for f^n we get $\rho(f^n(x), f^m(x)) \leq \frac{K^n + K^m}{1-K}(\rho(x, f(x)))$. 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. ∎

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. Immediate from the above. ∎

▷ Exercise 2. 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 ϵ 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:

Stopping Rule. *If* $d = \rho(x, f(x))$ *and* $N > (\log(\epsilon) + \log(1-K) - \log(d))/\log(K)$ *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 will 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 3. 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.

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 $\sigma(t) = p + t(q - p)$ 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 non-negative 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.)

A final remark. The standard proof of the Contraction Principle goes by comparing $\rho(f^n(p), f^{n+1}(p))$ with the geometric series K^n. While it is only slightly longer, it seems less natural and less elegant to me. I "discovered" the Fundamental Inequality, and the above proof, many years ago while trying to extend the Contraction Principle to a more general framework, but it is so obvious that I have no doubt that it has been discovered many times both before and after I noticed it.

1.2 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 that 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 δ is sufficiently small, then f maps $X = N_\delta(p)$ into itself and is a contraction on X. Of course we choose δ 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 δ 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's 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 zero's of a C^2 map $G : R^n \to R^n$ is relatively straightforward. Let $x_0 \in R^n$ 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 non-singular, and hence that DG_x is non-singular 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 $F : N_\delta(p) \to R^n$ 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 δ sufficiently small $F : N_\delta(p) \to N_\delta(p)$ is a contraction.

1.3 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 δ is sufficiently small, then f restricted to $X = N_\delta = \{v \in V | \, \|v\| \leq \delta\}$ is a contraction mapping of N_δ to itself provided $\|y\| < \delta/2$. By the Banach Contraction Principle it then follows that g maps N_δ 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_δ provided δ is sufficiently small. Since N_δ is convex, by a remark above this proves that $1/2$ is a Lipschitz bound for f in N_δ, and in particular that f restricted to N_δ is a contraction. Thus it only remains to show that f maps N_δ into itself provided $\|y\| < \delta/2$. That is, we must show that if $\|x\| \leq \delta$ then also $\|f(x)\| \leq \delta$. But since $f(0) = y$,

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

$\triangleright$ Exercise 4. 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 $\|x_1 - x_2\| \leq K \|y_1 - y_2\|$. Prove this with $K = 2$, using the fact that $h(y_i) = x_i$ is equivalent to $f_{y_i}(x_i) = x_i$ and that $1/2$ is a Lipschitz constant for $h = I - g$.

1.4 The Existence and Uniqueness Theorem for ODE.

Let V be an orthogonal vector space and let $F : V \times R \to V$ be a C^1 "time-dependent vector field" on V. If $I = [a, b]$ is an interval containing zero, then we call a C^1 curve $\sigma : I \to V$ a solution curve of the vector field F (or a solution of the ODE $dx/dt = F(x, t)$)

if $\sigma'(t) \equiv F(\sigma(t), t)$, and in this case we call the point $v_0 = \sigma(0)$ the initial condition of the solution.

Let $C(I, V)$ denote the space of all continuous maps $\sigma : I \to V$, and define a distance function on $C(I, V)$ by $\rho(\sigma_1, \sigma_2) = \max_{t \in I} \|\sigma_1(t) - \sigma_2(t)\|$. It is not hard to show that $C(I, V)$ is a complete metric space. (This just amounts to the theorem that a uniform limit of continuous functions is continuous.) Define for each v_0 in V a map $f = f_{v_0} : C(I, V) \to C(I, V)$ by $f(\sigma)(t) = v_0 + \int_0^t F(\sigma(s), s)\, ds$. We note that clearly $f(\sigma)(0) = v_0$ while, by the Fundamental Theorem of Integral Calculus $d/dt f(\sigma)(t) = F(\sigma(t), t)$. It follows that if σ is a fixed point of f then it is a solution curve of F with initial condition v_0, and the converse is just as obvious. It is therefore natural to try to find a solution curve of F with initial condition v_0 by starting with the constant path $\sigma_0(t) \equiv v_0$ and then iterating f on σ_0.

$\triangleright$ Exercise 5. Try this with the (time-independent) ODE on R given by $dx/dt = x$ using the initial condition $x(0) = 1$.

We will now see that given any point p in V there is an $\epsilon > 0$ and a $\delta > 0$ such that if we define $I = [-\delta, \delta]$, then for any $v_0 \in V$ with $\|v_0 - p\| < \epsilon$ there is a unique solution curve of F $\sigma : I \to V$ with initial condition v_0. This is called the Local Existence And Uniqueness Theorem for C^1 ODE.

Given any $\epsilon > 0$, using the technique explained earlier we can find a Lipschitz constant M for F restricted to the set of $(x, t) \in V \times R$ such that $\|x - p\| \le 2\epsilon$ and $|t| \le \epsilon$. Let B be the maximum value of $F(x, t)$ on this same set. Now choose $\delta > 0$ so that $K = M\delta < 1$ and $B\delta < \epsilon$, and define X to be the set of σ 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 desired result will follow from the Banach Contraction Principle if we can show that for $\|v_0\| < \epsilon$, f_{v_0} maps X into itself and has K as a Lipschitz bound.

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

$$\begin{aligned}
\|f_{v_0}(\sigma_1)(t) - f_{v_0}(\sigma_2)(t)\| &\le \int_0^t \|F(\sigma_1(s), s) - F(\sigma_2(s), s)\|\, ds \\
&\le \int_0^t M \|\sigma_1(s) - \sigma_2(s)\|\, ds \\
&\le \int_0^t M\rho(\sigma_1, \sigma_2)\, ds \\
&\le \delta M \rho(\sigma_1, \sigma_2) \le K\rho(\sigma_1, \sigma_2).
\end{aligned}$$

and it follows that $\rho(f_{v_0}(\sigma_1), f_{v_0}(\sigma_2) \le K\rho(\sigma_1, \sigma_2)$.