Fixed-Point Methods for Root-finding Problems with Functions of Several Variables

Functions of Several Independent and Dependent Variables. In the last few classes, we have introduced four critical methods for root-finding of a function with one variable: the Bisection Method, Fixed-point iteration, the Newton-Raphson Method and the Secant Method. In practice, problems tend to be in dimensions more than one, and dimension-reduction techniques can be also highly nontrivial (yet rewarding). We must face the fact that sometimes, we are dealt with a system of **nonlinear** equations,

$$\begin{aligned} f_1(x_1,\dots,x_n) &= 0, \\ f_2(x_1,\dots,x_n) &= 0, \\ &.. \\ &.. \\ &.. \\ f_n(x_1,\dots,x_n) &= 0, \end{aligned}$$

where each function f_i can be thought of as a mapping

$$f_i : \mathbb{R}^n \to \mathbb{R}.$$

Now, one can put together the f_i also into a vector,

$$\mathbf{F}(x_1,x_2,\dots,x_n) = (f_1(x_1,\dots,x_n), f_2(x_1,\dots,x_n),\dots,f_n(x_1,\dots,x_n))$$

as each f_i gives an independent output. Thus, the vector-valued function $\mathbf{F}$ is a mapping

$$\mathbf{F} : \mathbb{R}^n \to \mathbb{R}^n.$$

The root-finding problem for $\mathbf{F}$ becomes

$$\mathbf{F}(\mathbf{x}) = \mathbf{0}.$$

We say $f_1, f_2, \dots, f_n$ are the **coordinate functions** of $\mathbf{F}$.

Example. An ugly set of equations such as

$$\begin{aligned} 3e^{-x_1} + \sin(x_2x_3) + \frac{1}{2} &= 0 \\ x_1^2 + 4x_2 + x_3^{1/3} &= 0 \\ \cosh(x_1x_2) + 6x_3 &= 0 \end{aligned}$$

can be condensed in the format

$$\begin{aligned} \mathbf{F}(\mathbf{x}) &= \mathbf{F}(x_1,x_2,x_3) \\ &= (f_1(x_1,x_2,x_3), f_2(x_1,x_2,x_3), f_3(x_1,x_2,x_3))^{\mathrm{T}} \\ &= \left(3e^{-x_1} + \sin(x_2x_3) + \frac{1}{2}, x_1^2 + 4x_2 + x_3^{1/3}, \cosh(x_1x_2) + 6x_3\right)^{\mathrm{T}}. \end{aligned}$$

Properties of $\mathbf{F}$, such as limits, continuity, differentiability, etc., require a proper metric. In 1D, this metric is $|x-y|$, namely, the distance between two points. In higher dimensions, the metric is no other than the vector norm (such as l^2-norm) $\|\mathbf{x}-\mathbf{y}\|$. In fact, these properties are well-defined independent of the choice of the norm.

Fixed Points of $\mathbf{F}(\mathbf{x})$. We say a function $\mathbf{G} : D \subset \mathbb{R}^n \to \mathbb{R}^n$ has a **fixed point** at $\mathbf{p} \in D$ if $\mathbf{G}(\mathbf{p}) = \mathbf{p}$.

Recall that in one dimension, in order for $g(x)$ to have a unique fixed point, we require that g maps an interval to itself, and that $|g'(x)| \le k < 1$ for some nonzero k. A similar theorem can be proved for $\mathbf{G}$.

Theorem. *Let* $D = \left\{(x_1,x_2,\dots,x_n)^{\mathrm{T}} : a_i \le x_i \le b_i, \quad i = 1,2,\dots,n\right\}$ *for some collection of constants* $a_1,a_2,\dots,a_n$ *and* $b_1,b_2,\dots,b_n$. *Suppose* $\mathbf{G}$ *is a continuous function from* $D \subset \mathbb{R}^n$ *into* $\mathbb{R}^n$ *with the property that*

(1) *(existence)* $\mathbf{G}(\mathbf{x}) \in D$ *whenever* $\mathbf{x} \in D$, *and*

(2) *(uniqueness) There exists a constant $K<1$ such that every coordinate function g_i satisfies,*

$$\left|\frac{\partial g_i(\mathbf{x})}{\partial x_j}\right| \leq \frac{K}{n}, \quad \mathbf{x}\in D,$$

for each $j=1,2,\ldots,n$.

Then $\mathbf{G}$ has a fixed point in D, which can be found by the iterative scheme

$$\mathbf{x}^{(k)} = \mathbf{G}\left(\mathbf{x}^{(k-1)}\right), \quad k\geq 1.$$

This sequence converges to $\mathbf{p}\in D$ and the approximation satisfies error estimates

$$\left\|\mathbf{x}^{(k)}-\mathbf{p}\right\|_\infty \leq \frac{K^k}{1-K}\left\|\mathbf{x}^{(1)}-\mathbf{x}^{(0)}\right\|_\infty .$$

Example. Consider the nonlinear system

$$x_1^2-10x_1+x_2^2+8=0,$$
$$x_1x_2^2+x_1-10x_2+8=0.$$

We turn this into a fixed point iteration using the vector-valued function $\mathbf{G}(x_1,x_2)=(g_1(x_1,x_2),g_2(x_1,x_2))$ by identifying its coordinate functions

$$g_1(x_1,x_2)=\frac{x_1^2+x_2^2+8}{10},$$
$$g_2(x_1,x_2)=\frac{x_1x_2^2+x_1+8}{10}.$$

We see that

$$x_1^2-10x_1+x_2^2+8=0 \iff x_1=g_1(x_1,x_2),$$
$$x_1x_2^2+x_1-10x_2+8=0 \iff x_x=g_2(x_1,x_2),$$

which turns the root-finding problem into a fixed-point finding $\mathbf{x}=\mathbf{G}(\mathbf{x})$.

By some preliminary analysis, we find that $\mathbf{G}:\left[0,\frac{3}{2}\right]^2\to\left[0,\frac{3}{2}\right]^2$. Indeed, we check the extremes – if $x_1=x_2=0$, we have $0<g_1(0,0)=4/5<3/2$, and $0<g_2(0,0)=4/5<3/2$, while if $x_1=x_2=3/2$, $0<g_1(3/2,3/2)=1.35<3/2$ and $0<g_2(3/2,3/2)=1.2875<1.5$. Thus, existence of a fixed point is guaranteed.

To establish uniqueness, we check the partial derivatives. There are four of them. For $(x_1,x_2)\in\left[0,\frac{3}{2}\right]^2$, we find

$$\frac{\partial g_1}{\partial x_1}=\frac{x_1}{5}\leq\frac{3}{10}=\frac{12}{40}$$
$$\frac{\partial g_1}{\partial x_2}=\frac{x_2}{5}\leq\frac{3}{10}=\frac{12}{40}$$
$$\frac{\partial g_2}{\partial x_1}=\frac{x_2^2+1}{10}\leq\frac{13}{40}$$
$$\frac{\partial g_2}{\partial x_2}=\frac{x_1x_2}{10}\leq\frac{9}{40}$$

So, we may choose $K=\frac{15}{20}<1$ (this can be chosen even more tightly) so that

$$\left|\frac{\partial g_i(\mathbf{x})}{\partial x_j}\right|\leq\frac{K}{n}=\frac{15/20}{2}=\frac{15}{40}.$$

This guarantees that the fixed point is unique.

Then, we perform a functional iteration. To initiate, let's choose $\mathbf{x}^{(0)}=\left(x_1^{(0)},x_2^{(0)}\right)=(1,1)$ (better choose something inside D). We find

$$\mathbf{x}^{(1)}=\mathbf{G}\left(\mathbf{x}^{(0)}\right)=(g_1(1,1),g_2(1,1))=(1,1).$$

Oops, we found the fixed point in one step. Very lucky. Here is a program that shows convergence if we start elsewhere in D.

Acceleration using Gauss-Seidel. The functional iteration is of the form

$$\mathbf{x}^{(k)} = \mathbf{G}\left(\mathbf{x}^{(k-1)}\right)$$

where we evaluate the coordinate functions one by one, i.e.,

$$\begin{aligned} x_1^{(k)} &= g_1\left(x_1^{(k-1)}, x_2^{(k-1)}\right), \\ x_2^{(k)} &= g_2\left(x_1^{(k-1)}, x_2^{(k-1)}\right). \end{aligned}$$

Borrowing from the improvement of **Gauss-Seidel** on **Jacobi**, why don't we do

$$\begin{aligned} x_1^{(k)} &= g_1\left(x_1^{(k-1)}, x_2^{(k-1)}\right), \\ x_2^{(k)} &= g_2\left(\boxed{x_1^{(k)}}, x_2^{(k-1)}\right) \end{aligned}$$

where we use the immediate update within the same iteration? Indeed, see the improvement in convergence.

Newton's Method in Higher Dimensions

Newton's method in one dimension has a nice geometric interpretation. We approximate the function locally by its tangent line, and thus also approximate the zero of the function by the zero of the tangent line. In essence, we rely on the linearization formula (first-order Taylor expansion of f)

$$L(x) = f(x_0) + f'(x_0)(x - x_0)$$

and we pose that $L(x_1) = 0$ which yields

$$0 = f(x_0) + f'(x_0)(x_1 - x_0).$$

This, in turn, yields the formula for Newton-Raphson iteration,

$$x_1 = x_0 - \frac{f(x_0)}{f'(x_0)}$$

which we push forward the index one by one.

What if $\mathbf{F} = \mathbf{F}(x_1, x_2 \ldots, x_n) = (f_1(x_1, \ldots, x_n), \ldots, f_n(x_1, \ldots, x_n))$, a vector-valued function? Can we motivate a Newton's method in higher dimensions using a similar linearization formula? Maybe we are going too fast. Let's first consider $f(\mathbf{x}) = f(x_1, x_2, \ldots, x_n)$, just a scalar-valued function but with multiple variables. Its first-order Taylor expansion, i.e., linearization, looks like

$$L(\mathbf{x}) = f\left(\mathbf{x}^{(0)}\right) + \left[\frac{\partial f\left(\mathbf{x}^{(0)}\right)}{\partial x_1}, \ldots, \frac{\partial f\left(\mathbf{x}^{(0)}\right)}{\partial x_n}\right]\left(\mathbf{x} - \mathbf{x}^{(0)}\right) = f\left(\mathbf{x}^{(0)}\right) + \nabla f\left(\mathbf{x}^{(0)}\right)\left(\mathbf{x} - \mathbf{x}^{(0)}\right).$$

Now, treat each coordinate function of $\mathbf{F}$ as f, we simply replace f by f_i and list them in a column vector, namely,

$$\begin{bmatrix} L_1(\mathbf{x}) \\ L_2(\mathbf{x}) \\ \cdot \\ \cdot \\ L_n(\mathbf{x}) \end{bmatrix} = \begin{bmatrix} f_1\left(\mathbf{x}^{(0)}\right) \\ f_2\left(\mathbf{x}^{(0)}\right) \\ \cdot \\ \cdot \\ f_n\left(\mathbf{x}^{(0)}\right) \end{bmatrix} + \begin{bmatrix} \frac{\partial f_1\left(\mathbf{x}^{(0)}\right)}{\partial x_1} & \frac{\partial f_1\left(\mathbf{x}^{(0)}\right)}{\partial x_2} & \cdot & \cdot & \frac{\partial f_1\left(\mathbf{x}^{(0)}\right)}{\partial x_n} \\ \frac{\partial f_2\left(\mathbf{x}^{(0)}\right)}{\partial x_1} & \frac{\partial f_2\left(\mathbf{x}^{(0)}\right)}{\partial x_2} & \cdot & \cdot & \frac{\partial f_2\left(\mathbf{x}^{(0)}\right)}{\partial x_n} \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \cdot & \cdot & \cdot & \cdot & \cdot \\ \frac{\partial f_n\left(\mathbf{x}^{(0)}\right)}{\partial x_1} & \frac{\partial f_n\left(\mathbf{x}^{(0)}\right)}{\partial x_2} & \cdot & \cdot & \frac{\partial f_n\left(\mathbf{x}^{(0)}\right)}{\partial x_n} \end{bmatrix} \left(\mathbf{x} - \mathbf{x}^{(0)}\right)$$

or more compactly,

$$\mathbf{L}(\mathbf{x}) = \mathbf{F}\left(\mathbf{x}^{(0)}\right) + \mathbf{J_F}(\mathbf{x})\left(\mathbf{x} - \mathbf{x}^{(0)}\right).$$

This matrix of partial derivatives is called the **Jacobian** of $\mathbf{F}$, $\mathbf{J_F}(\mathbf{x})$ in short. Now, supposing that some $\mathbf{x}^{(1)}$ gives $\mathbf{L}\left(\mathbf{x}^{(1)}\right) = 0$, we then have

$$\mathbf{F}\left(\mathbf{x}^{(0)}\right) + \mathbf{J_F}\left(\mathbf{x}^{(0)}\right)\left(\mathbf{x}^{(1)} - \mathbf{x}^{(0)}\right) = 0$$

which implies

$$\mathbf{J_F}\left(\mathbf{x}^{(0)}\right)\left(\mathbf{x}^{(1)}-\mathbf{x}^{(0)}\right)=-\mathbf{F}\left(\mathbf{x}^{(0)}\right)$$
$$\mathbf{x}^{(1)}-\mathbf{x}^{(0)}=-\left[\mathbf{J_F}\left(\mathbf{x}^{(0)}\right)\right]^{-1}\mathbf{F}\left(\mathbf{x}^{(0)}\right)$$

and ultimately

$$\boxed{\mathbf{x}^{(1)}=\mathbf{x}^{(0)}-\left[\mathbf{J_F}\left(\mathbf{x}^{(0)}\right)\right]^{-1}\mathbf{F}\left(\mathbf{x}^{(0)}\right)},$$

the ever so powerful, Newton-Raphson iteration in high dimension. It requires an initial guess of a vector $\mathbf{x}^{(0)}$.

Compare this to the one-dimensional analog,

$$x^{(1)}=x^{(0)}-\frac{f\left(x^{(0)}\right)}{f'\left(x^{(0)}\right)},$$

the only difference is that the division now is replaced by matrix inversion.

In practice, we actually stop at

$$\mathbf{J_F}\left(\mathbf{x}^{(0)}\right)\left(\mathbf{x}^{(1)}-\mathbf{x}^{(0)}\right)=-\mathbf{F}\left(\mathbf{x}^{(0)}\right)$$

because this is in the form of

$$A\boldsymbol{y}=\boldsymbol{b}$$

where

$$A=\mathbf{J_F}\left(\mathbf{x}^{(0)}\right),\quad \boldsymbol{y}=\mathbf{x}^{(1)}-\mathbf{x}^{(0)},\quad \boldsymbol{b}=-\mathbf{F}\left(\mathbf{x}^{(0)}\right)$$

while we know what $\mathbf{x}^{(0)}$ is – so solving the linear system for $\boldsymbol{y}$ here informs us the value of $\mathbf{x}^{(1)}$.