Thursday, 9/5/13 Math 471, Fall 2013, Section 001 Lecture 2

1 Rootfinding for scalar equations

Definition 1. The root of a function $f(x)$ is a number p satisfying $f(p) = 0$.

Example 1:

- The function $f_1(x) = x^2 - 3x + 2$ has roots at $p = 1, 2$.
- The function $f_2(x) = x^2 - 3$ has roots at $p = \pm\sqrt{3}$.

△

Question 1: Linear functions and quadratic functions have familiar formulas; how can we find the roots of a general function $f(x)$?

1.1 Bisection method

Text: section 2.1

Idea: Find an interval $[a, b]$ such that $f(a)$ and $f(b)$ have opposite sign. Then $f(x)$ has a root in $[a, b]$ by the Intermediate Value Theorem (Math 451 - advanced calculus). Perhaps we could repeatedly shrink the interval $[a_n, b_n]$ on which $f(a_n)$ and $f(b_n)$ have opposite sign, until we get close enough that the length of the interval is smaller than the accuracy we need...

△

Algorithm 1: Bisection method

Input : a, b
Output: $x_n \approx p$

$n = 0$, $a_0 = a$, $b_0 = b$
while *not told to stop* **do**
  $x_n = \dfrac{a_n + b_n}{2}$ `% estimate of root`
  if $f(x_n) \cdot f(a_n) < 0$ **then**
    $a_{n+1} = a_n, \quad b_{n+1} = x_n$
  else
    $a_{n+1} = x_n, \quad b_{n+1} = b_n$
  end
  $n = n + 1$
end
return $p = x_n$

Question 2: How accurate is the bisection method?

We want to derive error bounds for the method; in this case, absolute error is given by

$$E_A = |p - x_n| \,.$$

$$|p - x_n| \leq |b_n - a_n| = \frac{1}{2}\,|b_{n-1} - a_{n-1}| = \left(\frac{1}{2}\right)^2 |b_{n-2} - a_{n-2}| \leq \cdots \leq \frac{|b_0 - a_0|}{2^n}$$

Interpretation: At each step of the bisection method, we gain one bit of accuracy.

△

Example 2: How many steps are needed to ensure that the absolute error is less than 10^{-3}?

$$\frac{|b-a|}{2^n} \leq 10^{-3} \implies n \geq 10$$

△

Question 3: How do we stop the algorithm (line 2)? You may have found the root to close enough approximation at step n, if, for a given tolerance $\epsilon > 0$,

- $|b_n - a_n| < \epsilon$, or
- $|f(x_n)| < \epsilon$, or
- $n = n_{\max}$.

Beware! Examples do exist where any one of these methods may not give a good enough answer! (See page 64). Draw a picture, or make a plot.

△

1.2 Fixed point iteration

Text: section 2.3

Given a function $f(x)$ whose root we seek (i.e., we want to solve $f(x) = 0$), suppose there is a separate function $g(x)$ such that the statement $f(x) = 0$ is equivalent to $g(x) = x$.

Definition 2. The equation $g(x) = x$ defines an iterative method, and $g(x)$ is called the iterative function.
Suppose that a sequence $\{x_n\}$ is generated by an initial guess x_0, and that $\lim_{n\to\infty} x_n = x$. Then, if $g(x)$ is continuous,

$$\lim_{n\to\infty} x_n = \lim_{n\to\infty} g(x_n) = g(x).$$

Example 3: We consider the function $f(x) = x^2 - 3$, which has a root at $p = \sqrt{3} = 1.73205\ldots$.
We can rewrite $f(x) = 0$ in many ways, for example:

$$\begin{aligned} x = \frac{3}{x} &\implies g_1(x) = \frac{3}{x} \\ x = x - (x^2 - 3) &\implies g_2(x) = x - (x^2 - 3) \\ x = x - \frac{x^2 - 3}{2} &\implies g_3(x) = x - \frac{x^2 - 3}{2} \end{aligned}$$

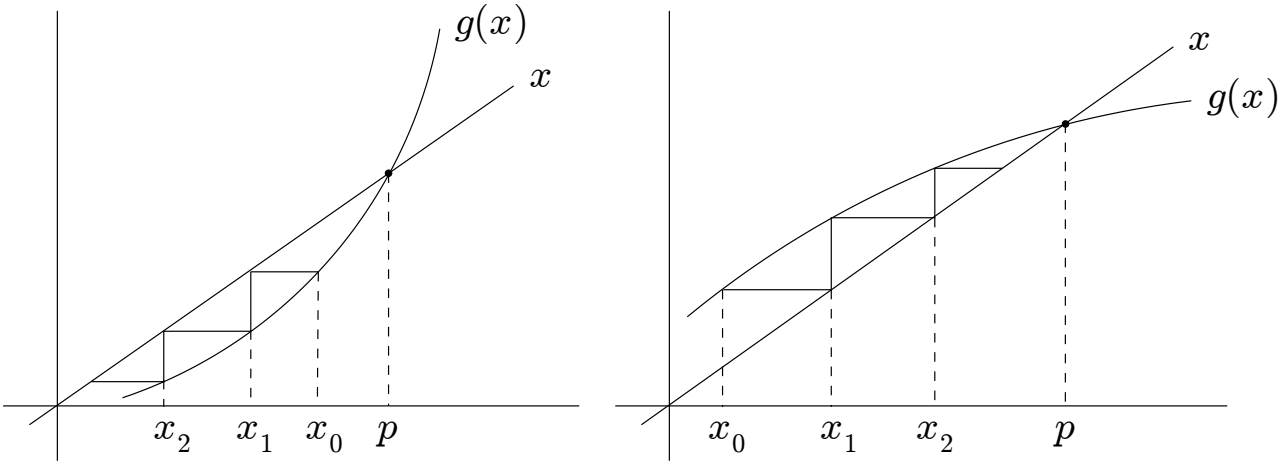
Next, we try to solve $g(x) = x$ by computing $x_{n+1} = g(x_n)$, starting with some initial guess x_0. This process is called fixed point iteration.

n	case 1 x_n	case 2 x_n	case 3 x_n
0	1.5	1.5	1.5
1	2	2.25	1.875
2	1.5	0.1875	1.6172
3	2	3.1523	1.8095
4	1.5	- 3.7849	1.6723
5	2	-15.1106	1.77399
6	1.5	-240.4409	1.70047
7	2	-58049.273	1.75467

Beware! Clearly, only case 3 appears to be converging...

△

Question 4: What determines whether a fixed point iteration converges or diverges? Can we use that information to design better iterative functions?
Let's take a look at two examples...



The first case diverges, while the second case converges... Why?

△

Theorem 3. *Suppose that the equation $x = g(x)$ has a root at $p = x$, and that in the interval*

$$I = \{x : |x - p| \leq \delta\}, \quad \delta > 0$$

$g'(x)$ exists and that $C = \max_{x \in I} |g'(x)|$. Then the fixed point iteration $x_{n+1} = g(x_n)$ converges if and only if $C < 1$, and

- *$x_n \in I$ for all $n = 0, 1, 2, \ldots$*
- *$\lim_{n \to \infty} x_n = p$*

- *p is the only fixed point of $x = g(x)$ in I.*

Proof. As an induction step, suppose that $x_n \in I$ and WLOG, assume $x_n < p$. By the mean value theorem (Math 451), there exists a point $\xi \in [x_n, p]$ such that

$$g'(\xi) = \frac{g(p) - g(x_n)}{p - x_n} \implies g(p) - g(x_n) = g'(\xi)(p - x_n).$$

Thus,

$$|p - x_{n+1}| = |g(p) - g(x_n)| = \left|g'(\xi)(p - x_n)\right| \leq \max_{x \in I} \left|g'(x)\right| |p - x_n| \leq C\delta.$$

Therefore $x_n \in I \implies x_{n+1} \in I$, and the part 1 is proved.
To show that $\lim_{n\to\infty} x_n = p$ we use the same logic,

$$|p - x_{n+1}| \leq C\,|p - x_n| \leq C^2\,|p - x_{n-1}| \leq C^3\,|p - x_{n-2}| \leq \cdots \leq C^n\,|p - x_0|$$

Since $C < 1$,

$$\begin{aligned} &\lim_{n\to\infty} |p - x_{n+1}| \leq \lim_{n\to\infty} C^n\,|p - x_0| = 0 \\ \implies &\lim_{n\to\infty} x_n = p. \end{aligned}$$

To prove that p is the only fixed point of $g(x)$ in I, we use proof by contradiction. Suppose that there exists another root $q \neq p$ in I.

$$p - q = g(p) - g(q) = g'(\xi)(p - q)$$

Therefore

$$|p - q| \leq C\,|p - q| < |p - q|\,;$$

the statement $|p - q| < |p - q|$ is clearly nonsense. We therefore conclude that the hypothesis $p \neq q$ must be false. □

Notes:

- We showed that $|p - x_n| \leq C\,|p - x_{n-1}|$; this is called linear convergence and C is called the asymptotic error constant.
- For x_0 sufficiently close to p, the asymptotic error constant $C \approx |g'(p)|$.

  Return to example: $f(x) = x^2 - 3$.

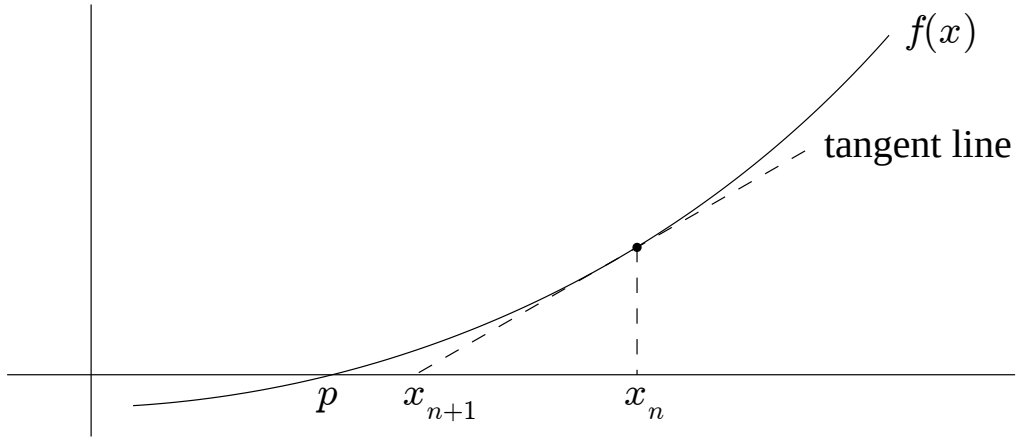
$$\begin{aligned} g_1(x) = \frac{3}{x} &\implies g_1'(x) = -\frac{3}{x^2} \implies \left|g_1'(p)\right| = 1 : \text{diverges} \\ g_2(x) = x - (x^2 - 3) &\implies g_2'(x) = 1 - 2x \implies \left|g_2'(p)\right| = 2.4641\ldots : \text{diverges} \\ g_3(x) = x - \frac{1}{2}(x^2 - 3) &\implies g_3'(x) = 1 - x \implies \left|g_3'(p)\right| = 0.732050\ldots : \text{converges} \end{aligned}$$

- The bisection method also converges linearly, with $C = \frac{1}{2}$.

1.3 Newton's method

Text: section 2.4

Idea: Local linear approximation. The x-intercept of the tangent line to $f(x)$ at $x = x_n$ will be near x-intercept of $f(x)$.



The tangent line passes through the point $\big(x_n, f(x_n)\big)$ with slope $f'(x_n)$, and is therefore defined the equation

$$y - f(x_n) = f'(x_n)(x - x_n).$$

We define x_{n+1} as the x-intercept of this line, so that

$$\begin{aligned} -f(x_n) &= f'(x_n)(x_{n+1} - x_n) \\ \implies x_{n+1} &= x_n - \frac{f(x_n)}{f'(x_n)} \end{aligned}$$

Convergence analysis: We recognize Newton's method as a fixed point iteration with iteration function $g(x) = x - \frac{f(x_n)}{f'(x_n)}$.
To guarantee that $g(x)$ is continuous near the root $x = p$, we assume that p is a simple root of $f(x)$, that is, $f(p) = 0$, but $f'(p) \neq 0$. Performance of Newton's method for roots of f with multiplicity greater than 1 are covered on pages 101-102.

$$\begin{aligned} g'(x) &= 1 - \frac{\big(f'(x)\big)^2 - f(x)f''(x)}{\big(f'(x)\big)^2} \\ &= 1 - \left(\frac{f'(x)}{f'(x)}\right)^2 - \frac{f(x)f''(x)}{\big(f'(x)\big)^2} \\ &= \frac{f(x)f''(x)}{\big(f'(x)\big)^2} \\ \implies g'(p) &= 0 \quad \text{since } f(p) = 0 \end{aligned}$$

Since $g(x)$ is continuous, there exists a small neighborhood near $x = p$ such that $|g(x)| \ll 1$. Since $g'(p) = 0$, we conclude that Newton's method must converge faster than linearly. To determine how fast, we examine the error in successive terms of the sequence $\{x_n\}$,

$$E_n = |p - x_n|\,, \qquad E_{n+1} = |p - x_{n+1}|\,.$$

Near $x = x_n$, the function $f(x)$ has the Taylor series expansion

$$f(x) = f(x_n) + f'(x_n)(x_n - x) + \frac{1}{2}f''(\xi)(x_n - x)^2 \quad \xi \in [x_n, p]$$

At the root p, we know that $f(p) = 0$, hence,

$$\begin{aligned}
0 &= f(x_n) + f'(x_n)(x_n - p) + \frac{1}{2}f''(\xi)(x_n - p)^2 \\
-\frac{f(x_n)}{f'(x_n)} &= x_n - p + \frac{f''(\xi)}{2f'(x_n)}(x_n - p)^2 \\
p - \underbrace{x_n - \frac{f(x_n)}{f'(x_n)}}_{x_{n+1}} &= \frac{f''(\xi)}{2f'(x_n)}(x_n - p)^2 \\
p - x_{n+1} &= \frac{f''(\xi)}{2f'(x_n)}(p - x_n)^2 \\
E_{n+1} &= \frac{f''(\xi)}{2f'(x_n)}E_n^2.
\end{aligned}$$

We have shown that Newton's method is quadratically convergent with asymptotic error constant

$$C = \frac{1}{2}\frac{|f''(p)|}{|f'(p)|}.$$

Example 4: The assumptions that lead to the ideal gas law, $PV = nRT$ are no longer valid at extreme pressure and temperature. Under such conditions, more processes must be included in the model equation, and we use the van der Waals equation instead

$$\left(P + \frac{n^2 a}{V^2}\right)(V - nb) = nRT.$$

In both equations P is pressure, V is volume, and T is temperature; R is the universal gas constant, $R = 0.08206$ atm·liter/(mole·K), and n is the number of moles.
In the van der Waals equation, the parameter a corrects the pressure due to forces between molecules that become significant at extreme pressure. The parameter b corrects the volume of the gas at very small scales where the volume of the actual molecules can no longer be ignored. Note that the limits $a \to 0$ and $b \to 0$ applied to van der Waals equation recover the ideal gas law.
For one mole of chlorine gas at a pressure of $P = 2$ atm and temperature of $T = 313$K, the parameters are

$$a = 6.29\frac{\text{atm} \cdot \text{L}^2}{\text{mole}^2} \qquad b = 0.0562\frac{\text{L}}{\text{mole}}.$$

What is the volume of the gas?

- We will produce our initial guess using the ideal gas law.

$$V_0 = \frac{RT}{P} = \frac{0.08206 \cdot 313}{2} = 12.842389999999998 \text{ atm}.$$

- Next we rewrite the equation to apply Newton's method.

$$\left(P + \frac{n^2 a}{V^2}\right)(V - nb) - nRT = 0 \implies f(V) = \left(P + \frac{n^2 a}{V^2}\right)(V - nb) - nRT$$
$$f'(V) = \left(P + \frac{n^2 a}{V^2}\right) + \left(\frac{-2n^2 a}{V^3}\right)(V - nb)$$

n	V_n
0	12.842389999999998
1	12.651154813406302
2	12.651099337119016

- We see that V_0 has 2 correct digits and V_1 has 5 correct digits.

  Question 5: How many correct digits does V_2 have? (hw)

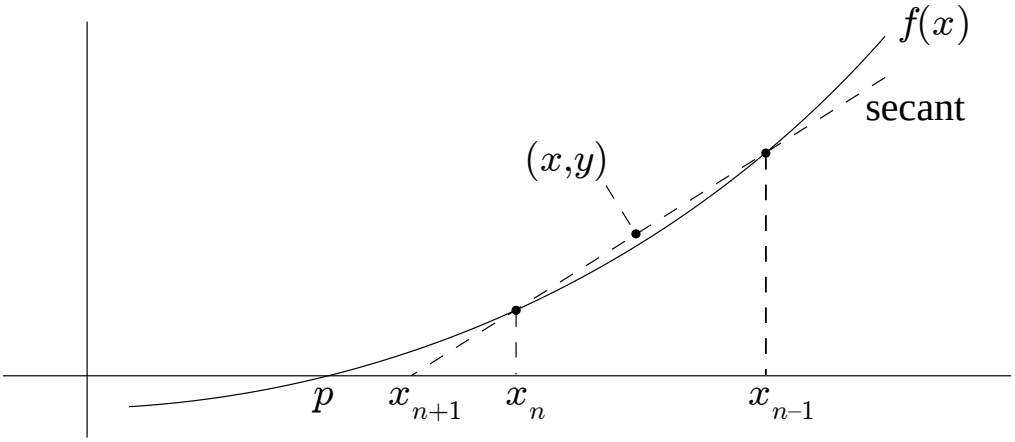
△

△

1.4 Secant method

Text: section 2.5

The quadratic convergence rate (to a simple root) of Newton's method is very attractive, but it comes at the cost of two function evaluations per step, which can be expensive computationally. Additionally, Newton's method requires knowledge of the derivative, $f'(x)$, which we may not have.
The construction of Newton's method can inform another technique, by replacing the tangent line with a secant line.



The secant line is defined by the points $\big(x_n, f(x_n)\big)$ and $\big(x_{n-1}, f(x_{n-1})\big)$, and is therefore given by the equation

$$y - f(x_n) = \frac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}(x - x_n).$$

We define x_{n+1} as the x-intercept of this line; plugging in $(x_{n+1}, 0)$ we find that,

$$x_{n+1} = x_n - \frac{f(x_n)}{\left(\frac{f(x_n)-f(x_{n-1})}{x_n - x_{n-1}}\right)} = x_n - \frac{f(x_n)(x_n - x_{n-1})}{f(x_n) - f(x_{n-1})}.$$

Notes:

- The secant method requires two steps, x_n and x_{n-1}, to determine x_{n+1}, and therefore requires two starting values, x_0 and x_1, to initialize.
- It only requires one new function evaluation $f(x_n)$ per step, and it does not require the derivative $f'(x)$.
- The secant method can also be derived by approximating the derivative in Newton's method as $f'(x_n) \approx \dfrac{f(x_n) - f(x_{n-1})}{x_n - x_{n-1}}$.
- It can be shown that $|p - x_n| \leq C\,|p - x_{n-1}|^{1.6}$ (pages 109-110), so the secant method converges faster than fixed-point iteration, but slower than Newton's method.

Summary:

method	rate of convergence	cost per step
bisection	linear, $k = \frac{1}{2}$	$f(x_n)$
fixed-point iteration	linear, $k = \lvert g'(p)\rvert$	$g(x_n)$
Newton	quadratic	$f(x_n),\ f'(x_n)$
secant	between linear and quadratic	$f(x_n)$

Beware! The bisection method is guaranteed to converge if the initial interval contains a root, but the other methods can be very sensitive to the choice of x_0, and may not converge at all.

Question 6: Each of these algorithms (bisection, fixed point, Newton, secant) have shown you how to find a single root. How would you solve a problem with multiple roots? (hw)

△

2 Rootfinding for nonlinear systems

Text: section 3.10

We have seen several methods that solve the problem of finding a root p of a scalar equation $f(x) = 0$, or the equivalent fixed point problem $x = g(x)$. Are the same ideas applicable to situations where the variables are vectors, i.e., can we use them to solve $F(\boldsymbol{x}) = 0$, or $\boldsymbol{x} = G(\boldsymbol{x})$, where F and G are general (possibly nonlinear) functions?

Notation: When speaking of vectors and systems of equations, subscripts denote the component number : $\boldsymbol{x} = [x_1, x_2, \ldots, x_n]$. Therefore, to represent sequences of vectors, we will use superscripts with parenthesis (to avoid confusion with exponents), so that the kth element of a vector sequence will be denoted by $\boldsymbol{x}^{(k)}$.

Definition 4. Let $\boldsymbol{x} \in \mathbb{R}^n$ and $F(\boldsymbol{x}) : \mathbb{R}^n \mapsto \mathbb{R}^n$. The Jacobian matrix of a differentiable vector function $F(\boldsymbol{x})$ is the matrix

$$J = \begin{bmatrix} \frac{\partial F_1}{\partial x_1} & \frac{\partial F_2}{\partial x_1} & \cdots & \frac{\partial F_n}{\partial x_1} \\ \frac{\partial F_1}{\partial x_2} & \frac{\partial F_2}{\partial x_2} & \cdots & \frac{\partial F_n}{\partial x_2} \\ \vdots & & \ddots & \vdots \\ \frac{\partial F_1}{\partial x_n} & \frac{\partial F_2}{\partial x_n} & \cdots & \frac{\partial F_n}{\partial x_n} \end{bmatrix},$$

i.e., it is the matrix whose i, j element is the derivative with respect to x_j of the ith component of $F(\boldsymbol{x})$.

1. The fixed-point iteration does generalize to systems of n nonlinear equations in n unknowns under a similar condition to the scalar case. If we are able to define a function $G(\boldsymbol{x})$ such that finding fixed points $\boldsymbol{x} = G(\boldsymbol{x})$ is equivalent to solving $F(\boldsymbol{x}) = 0$, **and** the absolute value of its Jacobian determinant

$$|J(\boldsymbol{x})| = \left\| \begin{bmatrix} \frac{\partial G_1}{\partial x_1} & \frac{\partial G_2}{\partial x_1} & \cdots & \frac{\partial G_n}{\partial x_1} \\ \frac{\partial G_1}{\partial x_2} & \frac{\partial G_2}{\partial x_2} & \cdots & \frac{\partial G_n}{\partial x_2} \\ \vdots & & \ddots & \vdots \\ \frac{\partial G_1}{\partial x_n} & \frac{\partial G_2}{\partial x_n} & \cdots & \frac{\partial G_n}{\partial x_n} \end{bmatrix} \right\| < 1,$$

   then the fixed point iteration of the system $\boldsymbol{x}^{(k+1)} = G(\boldsymbol{x}^{(k)})$ will converge.

   We will return to this idea later in the course.

2. Newton's method also generalizes into higher dimensional spaces. Like the scalar case, we can expand the vector function $F(\boldsymbol{x})$ in a Taylor series about $\boldsymbol{x}^{(k)}$,

$$F(\boldsymbol{x}) = F(\boldsymbol{x}^{(k)}) + F'(\boldsymbol{x}^{(k)})(\boldsymbol{x} - \boldsymbol{x}^{(k)}) + O\left(\left\|\boldsymbol{x} - \boldsymbol{x}^{(k)}\right\|^2\right),$$

   where $F'(\boldsymbol{x}^{(k)})$ is the Jacobian matrix of F, $J(\boldsymbol{x})$, evaluated at $\boldsymbol{x} = \boldsymbol{x}^{(k)}$.

   Using intuition we gained from our analysis of Newton's method for a scalar equation, we suspect that the subsequent term $\boldsymbol{x}^{(k+1)}$ satisfies $\boldsymbol{x}^{(k+1)} = C\left\|\boldsymbol{x}^{(k)} - \boldsymbol{p}\right\|^2$ for some constant C. Then $\boldsymbol{x}^{(k+1)}$ is much closer to zero than the current term $\boldsymbol{x}^{(k)}$. If we neglect terms quadratic and higher order terms in the Taylor expansion of $F(\boldsymbol{x})$, we are left with Newton's method for systems

$$F(\boldsymbol{x}^{(k)}) + J(\boldsymbol{x}^{(k)})(\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)}) = 0.$$

   This equation is a matrix equation of the form $A\boldsymbol{x} = b$, with $A = J(\boldsymbol{x}^{(k)})$, $\boldsymbol{x} = (\boldsymbol{x}^{(k+1)} - \boldsymbol{x}^{(k)})$, and $\boldsymbol{b} = -F(\boldsymbol{x}^{(k)})$.

   To solve this system, we have to find the matrix inverse of $J(\boldsymbol{x}^{(k)})$, thus, at each iteration of Newton's method, we have incur the computational cost of a matrix inversion. If we assume that $J(\boldsymbol{x})$ is invertible, then

$$\boldsymbol{x}^{(k+1)} = -\left(J(\boldsymbol{x}^{(k)})\right)^{-1} F(\boldsymbol{x}^{(k)}) + \boldsymbol{x}^{(k)}.$$

Matrix inversion via direct methods has an operation count of $O(n^3)$ for $n \times n$ matrices, unless there are special properties of the matrix that may be exploited. Indirect methods may also reduce the compuational cost of matrix inversion; both of these topic will be the subject of our next topic: linear algebra.

Note: When n is small, general direct methods are not hard to find. Recall : for 2×2 matrices,

$$A = \begin{bmatrix} a & b \\ c & d \end{bmatrix} \implies A^{-1} = \frac{1}{ad - bc} \begin{bmatrix} d & -b \\ -c & a \end{bmatrix}.$$

Example 5: page 141, chemical reactions

$\begin{matrix} 2A + B \rightleftharpoons C \\ A + D \rightleftharpoons C \end{matrix}$: reversible reactions for reactants A, B, D and product C

a_0, b_0, d_0 : initial concentrations (moles/liter) in chemical reactor (known)

c_1, c_2 : equilibrium concentrations of C produced by each reaction (unknown)

k_1, k_2 : equilibrium reaction constants (known)

These variables are related by the Law of Mass Action.

compound	equilibrium concentration
A	$a_0 - 2c_1 - c_2$
B	$b_0 - c_1$
C	$c_1 + c_2$
D	$d_0 - c_2$

$$\Rightarrow \begin{cases} k_1 = \dfrac{c_1 + c_2}{(a_0 - 2c_1 - c_2)^2 (b_0 - c_1)} \\ \\ k_2 = \dfrac{c_1 + c_2}{(a_0 - 2c_1 - c_2)(d_0 - c_2)} \end{cases}$$

Hence to find c_1, c_2 we need to solve a system of nonlinear equations.

$$\begin{aligned}
f(c_1, c_2) &= k_1 - (c_1 + c_2)(a_0 - 2c_1 - c_2)^{-2}(b_0 - c_1)^{-1} \\
g(c_1, c_2) &= k_2 - (c_1 + c_2)(a_0 - 2c_1 - c_2)^{-2}(d_0 - c_2)^{-1} \\
\frac{\partial f}{\partial c_1} &= -\frac{1}{(a_0 - 2c_1 - c_2)^2 (b_0 - c_1)} - \frac{4(c_1 + c_2)}{(a_0 - 2c_1 - c_2)^3 (b_0 - c_1)} - \frac{c_1 + c_2}{(a_0 - 2c_1 - c_2)^2 (b_0 - c_1)^2} \\
\frac{\partial f}{\partial c_2} &= -\frac{1}{(a_0 - 2c_1 - c_2)^2 (b_0 - c_1)} - \frac{2(c_1 + c_2)}{(a_0 - 2c_1 - c_2)^3 (b_0 - c_1)} \\
\frac{\partial g}{\partial c_1} &= -\frac{1}{(a_0 - 2c_1 - c_2)(b_0 - c_1)} - \frac{2(c_1 + c_2)}{(a_0 - 2c_1 - c_2)^2 (d_0 - c_2)} \\
\frac{\partial g}{\partial c_2} &= -\frac{1}{(a_0 - 2c_1 - c_2)(b_0 - c_1)} - \frac{c_1 + c_2}{(a_0 - 2c_1 - c_2)^2 (d_0 - c_2)} - \frac{c_1 + c_2}{(a_0 - 2c_1 - c_2)(d_0 - c_2)^2}
\end{aligned}$$

△