8.1 Root-finding

In this section we move from linear system to nonlinear problems. Our basic question is the following rootfinding problem:

> Given a continuous scalar function f of a scalar variable, find a real number r such that $f(r) = 0$.

Later we will extend the study to vector value functions, which naturally includes linear systems as a special case: $f(x) = Ax - b$: $\mathbb{R}^n \to \mathbb{R}^m$.

Let us consider the conditioning and error of finding the root. For simplicity, let us assume that $f(x)$ has continuous derivative. Suppose $f(r) = 0$ but we obtain $\widetilde{r}$ by the algorithm. We also assume that the error is from the input, that is, we have perturbed function $\widetilde{f}(x) = f(x) + \epsilon$ which gives $\widetilde{f}(\widetilde{r}) = 0$. We want to compute the absolute conditional number

$$\kappa = \lim_{\epsilon \to 0} \frac{|\widetilde{r} - r|}{|\widetilde{f} - f|},$$

namely, how the solution changes with the perturbation in function value. Denote $\delta = \widetilde{r} - r$. Using Taylor expansion, we have

$$0 = \widetilde{f}(\widetilde{r}) = f(r + \delta) + \epsilon \approx f(r) + f'(r)\delta + \epsilon + \mathcal{O}(\delta^2).$$

Therefore, we have $\kappa = |f'(x)|^{-1}$.

Unlike linear system, we can not obtain the exact solution of nonlinear equation in general case. One popular approach is the fixed point iteration, which transformed $f(x) = 0$ into another equivalent problem of finding $g(x) = x$. One typical example is $g(x) = x - f(x)$. The fixed point iteration has the following form:

$$x^{k+1} = g(x^k). \tag{8.1}$$

Immediately, the following questions arise:

1. Does fixed point iteration converge to the root of an equation?
2. If the sequence converges, at what rate is the error term $\varepsilon_k = |r - x^k|$ diminishing?

We want to know how ε_{k+1} varies against ε_k. We have

$$\varepsilon_{k+1} + r = g(\varepsilon_k + r) = g(r) + g'(r)\varepsilon_k + \frac{1}{2}g''(r)\varepsilon_k^2 + \cdots,$$

Since $g(r) = r$, we have $\varepsilon_{k+1} = g'(r)\varepsilon_k + \mathcal{O}(\varepsilon_k^2)$. Suppose at the k-th step we know ε_k is small enough that ε_k^2 is ignorable, we have

$$\varepsilon_{k+1} \approx g'(r)\varepsilon_k. \tag{8.2}$$

Clearly, if $|g'(r)| > 1$, ε_{k+1} will grow rather than decrease. Hence the fixed point iteration may diverge. If $|g'(r)| = \sigma < 1$, then we should have $|\varepsilon_k| = |x_k - r| \approx C\sigma^k$. The error converges zero at linear rate. We say that the rate is linear because the precision $\log|\varepsilon_k| \approx k \log\sigma + \log C$ grows at linear rate. The linear rate is asymptotic in the sense that we have (8.2) when x^k is in a small neighborhood of r (ε_k is very small), which often occurs when k is suffciently large.

More generally, we consider the class of Lipschitz functions. We say that g is L-Lipschitz continuous for some $L > 0$ on the interval S if

$$|g(s) - g(t)| \le L|s - t|, \quad \forall s, t \in S.$$

Immediately, we see that $|g'(s)| \le L$ is a special case of L-Lipschitzness because $g(s) - g(t) = \int_s^t g'(x)dx$. If $L < 1$, $g(x)$ is called a contraction map, because it reduces the distance of points after mapping.

Consider for example the iteration $g(x) = x - f(x)$, then Lipschitzness means $|s - t - f(s) - f(t)| \le L|s - t|$, implying that

$$(1 - L)|s - t| \le |f(s) - f(t)| \le (1 + L)|s - t|.$$

Therefore, $f(x)$ changes at most linearly across the domain.

The following theorem guarantees global convergence of fixed-point iteration under some mild conditions.

Theorem 1. *Suppose $g(x)$ is L-Lipschitz with $L < 1$ on an interval S. Then S contains exactly one fixed point r of g. And if the sequence x^k are in S, then $|x^k - r| \le L^{k-1}|x_1 - r|$ for all $k > 1$.*

Proof. Suppose there are two fixed points $r_1 \neq r_2$. Then $|r_1 - r_2| = |g(r_1) - g(r_2)| \le L|r_1 - r_2|$, a contradiction. We show the existence of fixed point by showing $\{x^k\}$ is a Cauchy sequence. That is, for any ε, there exists k such that $|x^s - x^t| \le \varepsilon$ when $s \geq t \geq k$. Then Cauchy sequence gaurantees convergence to alimit in the complete space.

Specifically, For any k, we have

$$|x^k - x^{k-1}| \le L|x^{k-1} - x^{k-2}| \le \cdots \le L^{k-1}|x^1 - x^0|.$$

Moreover,

$$|x^s - x^t| \le \sum_{i=t+1}^{s} |s^i - s^{i-1}| \le \sum_{i=t+1}^{s} L^{i-k}|x^k - x^{k-1}| \le \frac{L^{t+1-k}(1 - L^{s-t-1})}{1 - L}|x^k - x^{k-1}| \le \frac{L^k|x^1 - x^0|}{1 - L}.$$

It suffices to take $k \approx \log\big((1 - L)\varepsilon/|x^1 - x^0|\big)$.

Let r be the limit point, then it is easy to show $|x - r| \le L^{k-1}|x_1 - r|$. □

8.2 Newton method

Stronger convergence result can be guaranteed if we use more information about $f(x)$. Newton method uses the derivative $f'(x)$ as well as the function value $f(x)$. It then uses a linear approximation $l(x; x^k) = f(x^k) + f'(x^k)(x - x^k)$ to promote next move: Find x such that $l(x; x^k) = 0$. Clearly, the solution is

$$x^{k+1} = x^k - \frac{f(x^k)}{f'(x^k)}.$$

Next we analyze the convergence of Newton's method. Denote $\varepsilon_k = r - x^k$. We have

$$\varepsilon_{k+1} = \varepsilon_k + \frac{f(r-\varepsilon^k)}{f'(r-\varepsilon^k)} = \varepsilon_k + \frac{0 - \varepsilon_k f'(r) + \frac{1}{2}\varepsilon_k^2 f''(r) + \mathcal{O}(\varepsilon_k^3)}{f'(r) - \varepsilon_k f''(r) + \mathcal{O}(\varepsilon_k^2)}.$$

It then follows that

$$\varepsilon_{k+1} \approx \varepsilon_k + \frac{-\varepsilon_k f'(r) + \frac{1}{2}\varepsilon_k^2 f''(r)}{f'(r) - \varepsilon_k f''(r)} = -\frac{\frac{1}{2}\varepsilon_k^2 f''(r)}{f'(r) - \varepsilon_k f''(r)} \approx -\frac{1}{2}\frac{f''(r)}{f'(r)}\varepsilon_k^2.$$

Then each Newton step doubles the accuracy:

$$\log|\varepsilon_{k+1}| \approx 2\log|\varepsilon_k| + \text{const}.$$

8.3 Secant method

In many scenarios, it is difficult to compute derivative f' directly. In the secant method, one finds the root of the linear approximation through the two most recent root estimates. That is, given previous approximations $x_1 \ldots, x_k$, we denote

$$l(x, x^k) = f(x^k) + \frac{f(x^k) - f(x^{k-1})}{x^k - x^{k-1}}(x - x^k).$$

The next iterate x^{k+1} is the root of $l(x; x^k) = 0$:

$$x^{k+1} = x^k - \frac{f(x^k)(x^k - x^{k-1})}{f(x^k) - f(x^{k-1})}.$$

By using $\varepsilon_k = r - x^k$, we have

$$\frac{f(x^k) - f(x^{k-1})}{x^k - x^{k-1}} \approx \frac{(\varepsilon_{k-1} - \varepsilon_k)\, f'(r) + \frac{1}{2}\left(\varepsilon_k^2 - \varepsilon_{k-1}^2\right) f''(r)}{\varepsilon^{k-1} - \varepsilon^k} = f'(r) - \frac{1}{2}\left(\varepsilon_{k-1} + \varepsilon_k\right) f''(r)$$

and

$$\varepsilon_{k+1} \approx \varepsilon_k + \frac{\left[-\varepsilon_k f'(r) + \frac{1}{2}\varepsilon_k^2 f''(r)\right]}{f'(r) - \frac{1}{2}\left(\varepsilon_{k-1} + \varepsilon_k\right) f''(r)} \approx -\frac{1}{2}\frac{f''(r)}{f'(r)}\varepsilon_k\varepsilon_{k-1}.$$

Suppose $\varepsilon_{k+1} = c\varepsilon_k^\alpha$, $\forall k$. Then we have

$$\left(\varepsilon_{k-1}^\alpha\right)^\alpha \approx \varepsilon_{k-1}^\alpha \varepsilon_{k-1}.$$

The solution is $\alpha = \frac{1+\sqrt{5}}{2} \approx 1.618$. The convergence rate is superlinear, slower than that of Newton method but better than the linear rate. However, in each round secant method needs only one new function call while Newton method needs two ($f(x^k)$ and $f'(x^k)$). In Newton method, we have

$$\varepsilon_{k+1} = \varepsilon_k^2 = \left(\varepsilon_k^{\sqrt{2}}\right)^{\sqrt{2}}.$$

Hence each oracle effectively increases $\sqrt{2}$ bit accuracy. Therefore, the effective convergence rate of Newton is superlinear with exponent $\sqrt{2} < (1+\sqrt{5})/2$.

8.4 Nonlinear systems

We can also extend Newton's method to solving nonlinear equations. Consider $f : \mathbb{R}^n \to \mathbb{R}^n$. For $x \in \mathbb{R}^n$

$$f(x) = \begin{pmatrix} f_1(x) \\ f_2(x) \\ \vdots \\ f_n(x) \end{pmatrix}.$$

Define the Jacobian matrix

$$J(x) = \begin{bmatrix} \frac{\partial f_1}{\partial x_1} & \frac{\partial f_1}{\partial x_2} & \cdots & \frac{\partial f_1}{\partial x_n} \\ \frac{\partial f_2}{\partial x_1} & & & \\ \vdots & & & \vdots \\ \frac{\partial f_n}{\partial x_1} & \frac{\partial f_n}{\partial x_2} & \cdots & \frac{\partial f_n}{\partial x_n} \end{bmatrix},$$

and the linear map $L(x, x^k) = f(x^k) + J(x^k)(x - x^k)$. By requiring $L(x, x^k) = 0$, we have $x^{k+1} = x^k - \left[J(x^k)\right]^{-1} f(x^k)$. Hence in each Newton step, we need to solve a linear system

$$J_k s^k = -f(x^k)$$

and update $x^{k+1} = x^k + s^k$.

We can also extend secant method to high dimensional case by using the finite difference to approximate Jacobian. But then we have to estimate $\frac{\partial f_j}{\partial x_i}$ for each pair of (i, j):

$$\frac{\partial f_j}{\partial x_i} \approx \frac{f_j(x + \delta e_i) - f_j(x)}{\delta}.$$

This can be quite time-consuming when n is large.

The Quasi-Newton methods find much more convenient way to compute J_k. The intuition is that the Jacobian matrices look similar when the algorithm converges, hence there may be some cheap way of updating J_{k+1} based on the existing J_k, instead of computing J_k independently in each round.

Let A^k be the approximation of $J(x^k)$. Analogous to the secant method, we want the new approximate Jacobian A_{k+1} to satisfy:

$$A^{k+1}(x^{k+1} - x^k) = f(x^{k+1}) - f(x^k). \tag{8.3}$$

But this is only one constraint, not enough to determine A^{k+1}. Therefore, we impose rank-one update of A^k such that (8.3) is satisfied. In Broyden's version of Quasi-Newton method, the update has the form:

$$A^{k+1} = A^k + \frac{1}{\|s^k\|^2} \left(f(x^{k+1}) - f(x^k) - A^k s^k\right) \left(s^k\right)^T.$$

Overdetermined nonlinear system It is possible to extend our study to nonlinear system of $f : \mathbb{R}^n \to \mathbb{R}^m$ with $m > n$. To this end, we consider the nonlinear least squares problem:

$$\min_x \|f(x)\|_2.$$

Similarly, by linear approximation $L(x, x^k) = f(x^k) + J(x^k)(x - x^k)$ we solve a sequence of least squares problems:

$$s^k = \operatorname*{argmin}_s \left\|f(x^k) + J(x^k)s\right\|$$

and $x^{k+1} = x^k + s^k$.