Newton's method: Basic Idea

Objective:

$$\min_{x\in\mathbb{R}^n} f(x), \quad f:\mathbb{R}^n \to \mathbb{R}$$

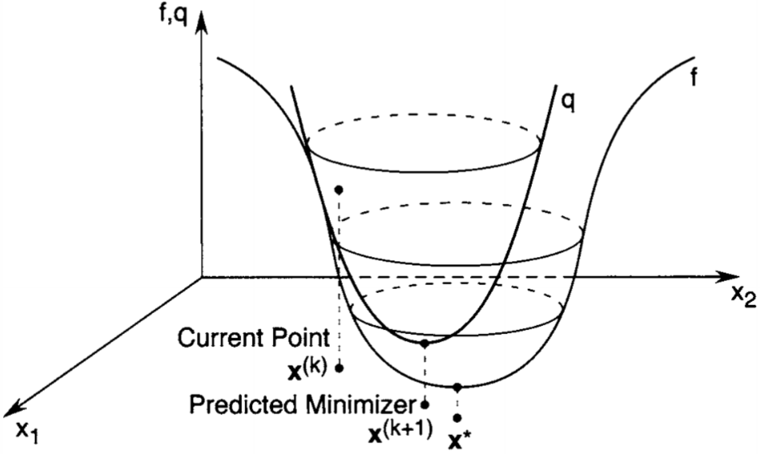
Given the current point $x^{(k)}$

- construct a quadratic function (known as the quadratic approximation;using Tayor's approximation) to the objective function that matches the value and both the first and second derivatives at $x^{(k)}$
- minimize the quadratic function instead of the original objective function
- set the minimizer as $x^{(k+1)}$

Note: a new quadratic approximation will be constructed at $x^{(k+1)}$

Special case: the objective is quadratic, the approximation is exact and the method returns a solution in one step.

Geometric Illustration



- Assumption: function $f \in \mathcal{C}^2$, i.e., twice continuously differentiable
- Apply Taylor's expansion, keep first three terms, drop terms of order ≥ 3

$$f(\boldsymbol{x}) \approx q(\boldsymbol{x}) := f(\boldsymbol{x}^{(k)}) + \boldsymbol{g}^{(k)T}(\boldsymbol{x} - \boldsymbol{x}^{(k)}) + \frac{1}{2}(\boldsymbol{x} - \boldsymbol{x}^{(k)})^T \boldsymbol{F}(\boldsymbol{x}^{(k)})(\boldsymbol{x} - \boldsymbol{x}^{(k)})$$

where

  - $\boldsymbol{g}^{(k)} := \nabla \boldsymbol{f}(\boldsymbol{x}^{(k)})$ is the gradient at $\boldsymbol{x}^{(k)}$
  - $\boldsymbol{F}(\boldsymbol{x}^{(k)}) := \nabla^2 \boldsymbol{f}(\boldsymbol{x}^{(k)})$ is the Hessian at $\boldsymbol{x}^{(k)}$

- Minimizing $q(\boldsymbol{x})$ by apply the first-order necessary condition:

$$\mathbf{0} = \nabla q(\boldsymbol{x}) = \boldsymbol{g}^{(k)} + \boldsymbol{F}(\boldsymbol{x}^{(k)})(\boldsymbol{x} - \boldsymbol{x}^{(k)}).$$

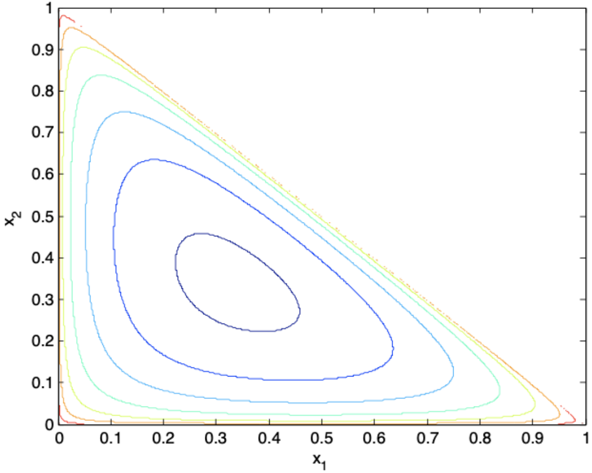
- If $\boldsymbol{F}(\boldsymbol{x}^{(k)}) \succ 0$ (positive definite), then q achieves its unique minimizer at

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

We have $\mathbf{0} = \nabla q(\boldsymbol{x}^{(k+1)})$

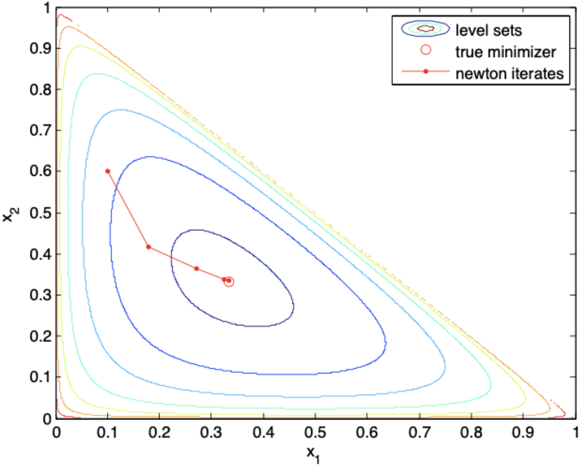
Example

$$f(x_1, x_2) = -\log(1 - x_1 - x_2) - \log(x_1) - \log(x_2)$$



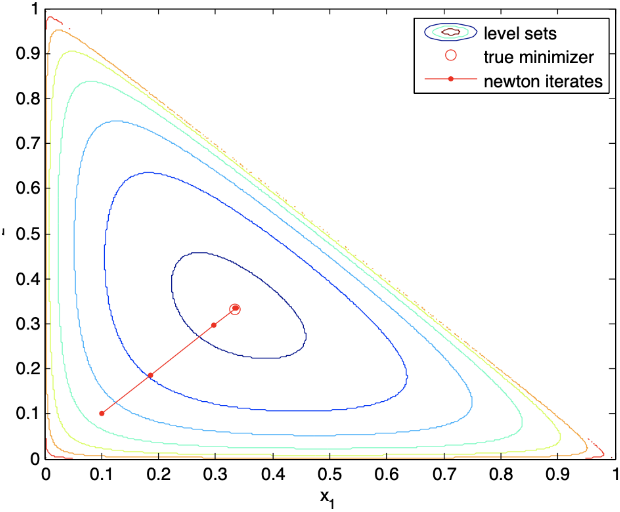
Example

Start Newton's method from $(\frac{1}{10}, \frac{6}{10})$



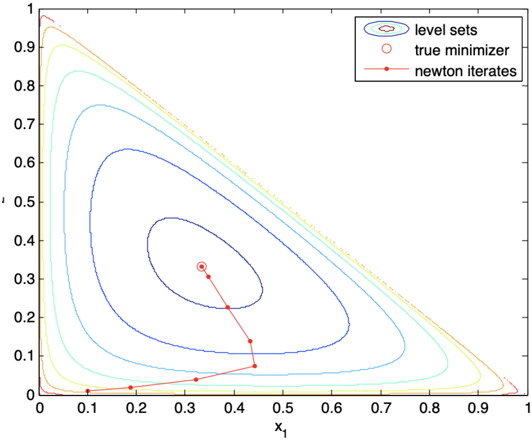
Example

Start Newton's method from $\left(\frac{1}{10}, \frac{1}{10}\right)$



Example

Start Newton's method from $\left(\frac{1}{100}, \frac{1}{100}\right)$



Quadratic function minimization

- The objective function

$$f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{x}^T\boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}^T\boldsymbol{x}$$

Assumption: $\boldsymbol{Q}$ is symmetric and invertible

$$\boldsymbol{g}(\boldsymbol{x}) = \boldsymbol{Q}\boldsymbol{x} - \boldsymbol{b}$$
$$\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{Q}.$$

- First-order optimality condition $\boldsymbol{g}(\boldsymbol{x}^*) = \boldsymbol{Q}\boldsymbol{x}^* - \boldsymbol{b} = \boldsymbol{0}$. So, $\boldsymbol{x}^* = \boldsymbol{Q}^{-1}\boldsymbol{b}$.
- Given any initial point $\boldsymbol{x}^{(0)}$, by Newton's method

$$\begin{aligned}
\boldsymbol{x}^{(1)} &= \boldsymbol{x}^{(0)} - \boldsymbol{F}(\boldsymbol{x}^{(0)})^{-1}\boldsymbol{g}^{(0)} \\
&= \boldsymbol{x}^{(0)} - \boldsymbol{Q}^{-1}(\boldsymbol{Q}\boldsymbol{x}^{(0)} - \boldsymbol{b}) \\
&= \boldsymbol{Q}^{-1}\boldsymbol{b} \\
&= \boldsymbol{x}^*.
\end{aligned}$$

The solution is obtained in one step.

How Fast Is Newton's Method?

Suppose $f \in C^3$ and $x^* \in \mathbb{R}^n$ is a point such that

$$\nabla f(x^*) = 0 \qquad F(x^*) \succ 0.$$

Then for all $x^{(0)}$ sufficiently close to x^*, Newton's method is well defined for all k, and there exists a $C > 0$ such that

$$\|x^{(j+1)} - x^*\| \leq C\|x^{(j)} - x^*\|^2, \quad j = k, k+1, k+2, \ldots$$

(This means that the order of convergence is two.)

Asymptotic rates of convergence

Suppose sequence $\{\boldsymbol{x}^k\}$ converges to $\bar{\boldsymbol{x}}$. Perform the ratio test

$$\lim_{k\to\infty} \frac{\|\boldsymbol{x}^{k+1} - \bar{\boldsymbol{x}}\|}{\|\boldsymbol{x}^k - \bar{\boldsymbol{x}}\|} = \mu.$$

- if $\mu = 1$, then $\{\boldsymbol{x}^k\}$ converges **sublinearly**.
- if $\mu \in (0,1)$, then $\{\boldsymbol{x}^k\}$ converges **linearly**;
- if $\mu = 0$, then $\{\boldsymbol{x}^k\}$ converges **superlinearly**;

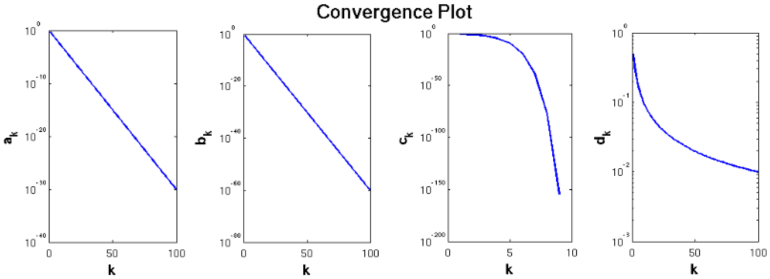
To distinguish superlinear rates of convergence, we check

$$\lim_{k\to\infty} \frac{\|\boldsymbol{x}^{k+1} - \bar{\boldsymbol{x}}\|}{\|\boldsymbol{x}^k - \bar{\boldsymbol{x}}\|^q} = \mu > 0$$

- if $q = 2$, it is **quadratic convergence**;
- if $q = 3$, it is **cubic convergence**;
- q can be non-integer, e.g., 1.618 for the secant method ...

Example: Linear, linear, superlinear (quadratic), sublinear

- $a_k = 1/2^k$
- $b_k = 1/4^{\lfloor k/2 \rfloor}$
- $c_k = 1/2^{2^k}$
- $d_k = 1/(k+1)$



"semilog$y$" plots (wikipedia)

When is Newton's Direction a Descent Direction?

At the kth iterate, if

$$F(x^{(k)}) \succ 0 \qquad g^{(k)} = \nabla f(x^{(k)}) \neq 0$$

then the search direction

$$d^{(k)} = -F(x^{(k)})^{-1} g^{(k)}$$

is a descent direction, that is, there exists $\bar{\alpha} > 0$ such that

$$f(x^{(k)} + \alpha d^{(k)}) < f(x^{(k)}), \qquad \forall \alpha \in (0, \bar{\alpha})$$

Two More Issues with Newton's Method

Indefinite Hessian:

- When the Hessian is not positive definite, the direction is not necessarily a descend direction.
- A simple solution is to use Levenberg-Marquardt approach!

Hessian evaluation:

- When the dimension n is large, obtaining $F(x^{(k)})$ can be computationally expensive
- Quasi-Newton method can be used to alleviate this difficulty!

Levenberg-Marquardt for Indefinite Hessian

If the Hessian $F(x^{(k)})$ is not positive definite, then the search direction

$$d^{(k)} = -F(x^{(k)})^{-1} g^{(k)}$$

may not point in a descent direction. A simple technique to ensure that the search direction is a descent direction is to used the so-called Levenberg-Marquardt modification to Newton's algorithm:

$$x^{(k+1)} = x^{(k)} - (F(x^{(k)}) + \mu_k I)^{-1} g^{(k)}$$

where $\mu_k \geq 0$.

Idea Underlying the Levenberg-Marquardt Modification

Let F be an *nxn* symmetric matrix but not be positive definite. The eigenvalues and eigenvectors of F are given by

$$\lambda_1, \lambda_2, \ldots, \lambda_n, \qquad v_1, v_2, \ldots, v_n \in \mathbb{R}^n$$

Note that all λ_i's are real, but not all positive (why?).

Now consider the matrix $G = F + \mu I, \mu \geq 0$. The eigenvalues of G are

$$\lambda_1 + \mu, \;\; \lambda_2 + \mu, \;\; \ldots, \;\; \lambda_n + \mu$$

$$Gv_i = (F + \mu I)v_i = Fv_i + \mu I v_i = \lambda_i v_i + \mu v_i = (\lambda_i + \mu)v_i$$

which shows that for all $i = 1, \ldots, n$ v_i is an eigenvector of G with eigenvalue $\lambda_i + \mu$. If μ is sufficiently large, then all eigenvalues of G are positive and G is positive definite.

In practice, we may start with a small value of μ_k, and then slowly increase it until we find that the iteration is descent, that is,

$$f(x^{(k+1)}) < f(x^{(k)}).$$

Gauss-Newton's Method

- Given functions $r_i : \mathbb{R}^n \to \mathbb{R},\ i = 1, \ldots, m$
- The goal is to find $\boldsymbol{x}^*$ so that $r_i(\boldsymbol{x}) = 0$ or $r_i(\boldsymbol{x}) \approx 0$ for all i.
- Consider the nonlinear least-squares problem

$$\underset{\boldsymbol{x}}{\text{minimize}} \ \frac{1}{2} \sum_{i=1}^{m} (r_i(\boldsymbol{x}))^2 .$$

- Define $\boldsymbol{r} = [r_1, \ldots, r_m]^T$. Then we have

$$\underset{\boldsymbol{x}}{\text{minimize}}\ f(\boldsymbol{x}) = \frac{1}{2}\boldsymbol{r}(\boldsymbol{x})^T \boldsymbol{r}(\boldsymbol{x}).$$

- The gradient $\nabla f(\boldsymbol{x})$ is formed by components

$$(\nabla f(\boldsymbol{x}))_j = \frac{\partial f}{\partial x_j}(\boldsymbol{x}) = \sum_{i=1}^{m} r_i(\boldsymbol{x}) \frac{\partial r_i}{\partial x_j}(\boldsymbol{x})$$

- Define the Jacobian of $\boldsymbol{r}$

$$\boldsymbol{J}(\boldsymbol{x}) = \begin{bmatrix} \frac{\partial r_1}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial r_i}{\partial x_n}(\boldsymbol{x}) \\ & \cdots & \\ \frac{\partial r_m}{\partial x_1}(\boldsymbol{x}) & \cdots & \frac{\partial r_m}{\partial x_n}(\boldsymbol{x}) \end{bmatrix}$$

  Then, we have

$$\nabla f(\boldsymbol{x}) = \boldsymbol{J}(\boldsymbol{x})^T \boldsymbol{r}(\boldsymbol{x})$$

- The Hessian $\boldsymbol{F}(\boldsymbol{x})$ is symmetric matrix. Its (k, j)th component is

$$\frac{\partial^2 f}{\partial x_k \partial x_j} = \frac{\partial}{\partial x_k}\left(\sum_{i=1}^{m} r_i(\boldsymbol{x})\frac{\partial r_i}{\partial x_j}(\boldsymbol{x})\right)$$
$$= \sum_{i=1}^{m}\left(\frac{\partial r_i}{\partial x_k}(\boldsymbol{x})\frac{\partial r_i}{\partial x_j}(\boldsymbol{x}) + r_i(\boldsymbol{x})\frac{\partial^2 r_i}{\partial x_k \partial x_j}(\boldsymbol{x})\right)$$

- Let $\boldsymbol{S}(\boldsymbol{x})$ be formed by (k, j)th components

$$\sum_{i=1}^{m} r_i(\boldsymbol{x})\frac{\partial^2 r_i}{\partial x_k \partial x_j}(\boldsymbol{x})$$

- Then, we have $\boldsymbol{F}(\boldsymbol{x}) = \boldsymbol{J}(\boldsymbol{x})^T\boldsymbol{J}(\boldsymbol{x}) + \boldsymbol{S}(\boldsymbol{x})$

- Therefore, Newton's method has the iteration

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \underbrace{\left(\boldsymbol{J}(\boldsymbol{x})^T\boldsymbol{J}(\boldsymbol{x}) + \boldsymbol{S}(\boldsymbol{x})\right)^{-1}}_{\boldsymbol{F}(\boldsymbol{x})^{-1}}\underbrace{\boldsymbol{J}(\boldsymbol{x})^T\boldsymbol{r}(\boldsymbol{x})}_{\nabla f(\boldsymbol{x})}$$

The Gauss-Newton method

- When the matrix $\boldsymbol{S}(\boldsymbol{x})$ is ignored in some applications to save computation, we arrive at the Gauss-Newton method

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - \underbrace{(\boldsymbol{J}(\boldsymbol{x})^T \boldsymbol{J}(\boldsymbol{x}))^{-1}}_{(\boldsymbol{F}(\boldsymbol{x})-\boldsymbol{S}(\boldsymbol{x}))^{-1}} \underbrace{\boldsymbol{J}(\boldsymbol{x})^T \boldsymbol{r}(\boldsymbol{x})}_{\nabla f(\boldsymbol{x})}$$

- A potential problem is that $\boldsymbol{J}(\boldsymbol{x})^T \boldsymbol{J}(\boldsymbol{x}) \not\succ \mathbf{0}$ and $f(\boldsymbol{x}^{(k+1)}) \geq f(\boldsymbol{x}^{(k)})$.

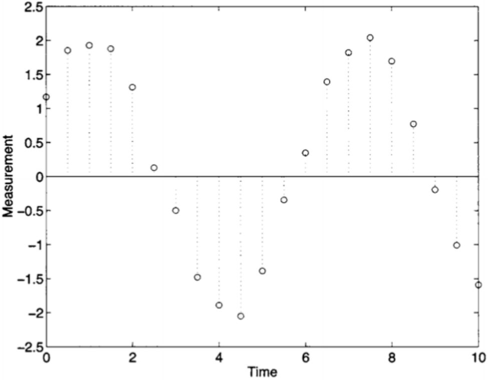
  Fixes: line search, Levenberg-Marquardt, and Cholesky/Gill-Murray.

Example: nonlinear data-fitting

- Given a sinusoid

$$y = A\sin(\omega t + \phi)$$

- Determine parameters A, ω, and ϕ so that the sinusoid best fits the observed points: (t_i, y_i), $i = 1, \ldots, 21$.



- Let $\boldsymbol{x} := [A, \omega, \phi]^T$ and

$$r_i(\boldsymbol{x}) := y_i - A\sin(\omega t_i + \phi)$$

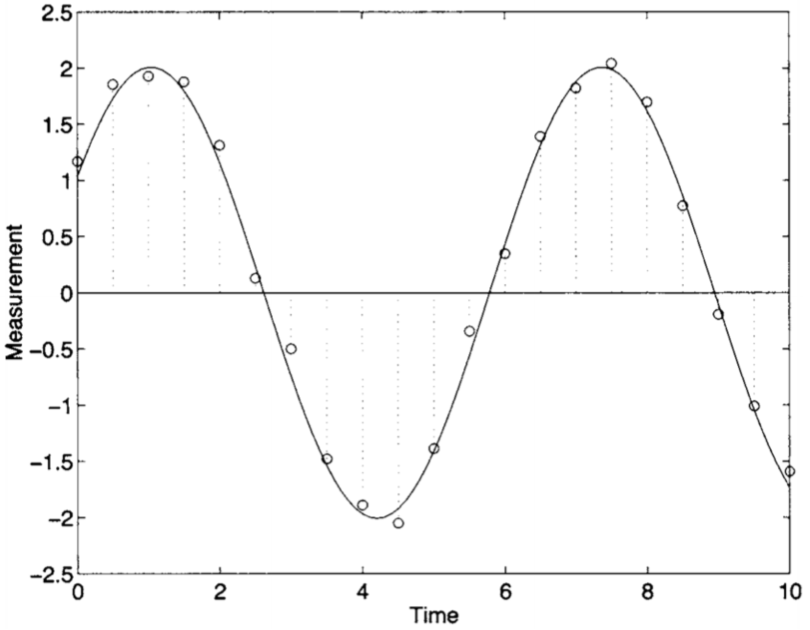
- Problem

$$\text{minimize} \sum_{i=1}^{21} (\underbrace{y_i - A\sin(\omega t_i + \phi)}_{r_i(\boldsymbol{x})})^2$$

- Derive $\boldsymbol{J}(\boldsymbol{x}) \in \mathbb{R}^{21\times 3}$ and apply the Gauss-Newton iteration

$$\boldsymbol{x}^{(k+1)} = \boldsymbol{x}^{(k)} - (\boldsymbol{J}(\boldsymbol{x})^T \boldsymbol{J}(\boldsymbol{x}))^{-1} \boldsymbol{J}(\boldsymbol{x})^T \boldsymbol{r}(\boldsymbol{x})$$

- Results: $A = 2.01,\ \omega = 0.992,\ \phi = 0.541.$


Measurement
Time
2.5
2
1.5
1
0.5
0
−0.5
−1
−1.5
−2
−2.5
0
2
4
6
8
10

Conclusions

Although Newton's method has many issues, such as

- the direction can be ascending if $\boldsymbol{F}(\boldsymbol{x}^{(k)}) \nsucc 0$
- may not ensure descent in general
- must start close to the solution,

Newton's method has the following strong properties:

- one-step solution for quadratic objective with an invertible $\boldsymbol{Q}$
- second-order convergence rate near the solution if $\boldsymbol{F}$ is Lipschtiz
- a number of modifications that address the issues.