Newton's methods

Instructor: Jin Zhang

Department of Mathematics
Southern University of Science and Technology
Spring 2024

Main features of Newton's method

- Uses both first derivatives (gradients) and second derivatives (Hessian)
- Based on local quadratic approximations to the objective function
- Requires a positive definite Hessian to work
- Converges very quickly near the solution (under conditions)
- Require a lot of work at each iteration:
  - forming the Hessian
  - inverting or factorizing the (approximate) Hessian

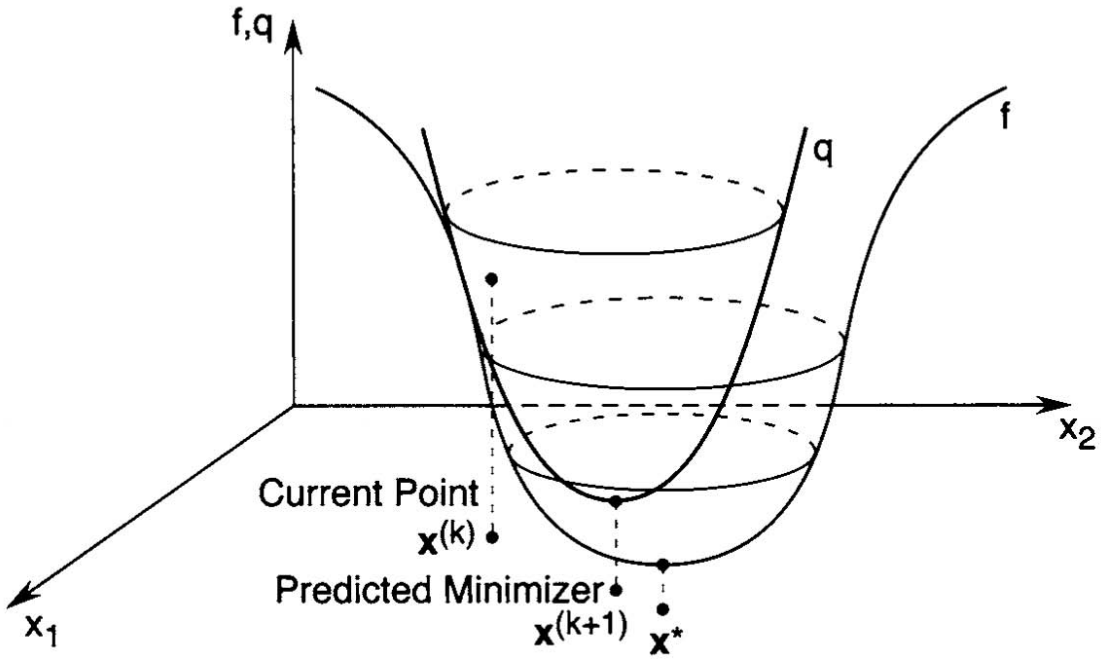
Basic idea

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

- construct a quadratic function (known as the quadratic 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


f,q
q
f
x_2
x_1
Current Point
$x^{(k)}$
Predicted Minimizer
$x^{(k+1)}$
x^*

Quadratic approximation

- 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

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

where

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

Generating the next point

- Minimizing $q(x)$ by apply the first-order necessary condition:

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

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

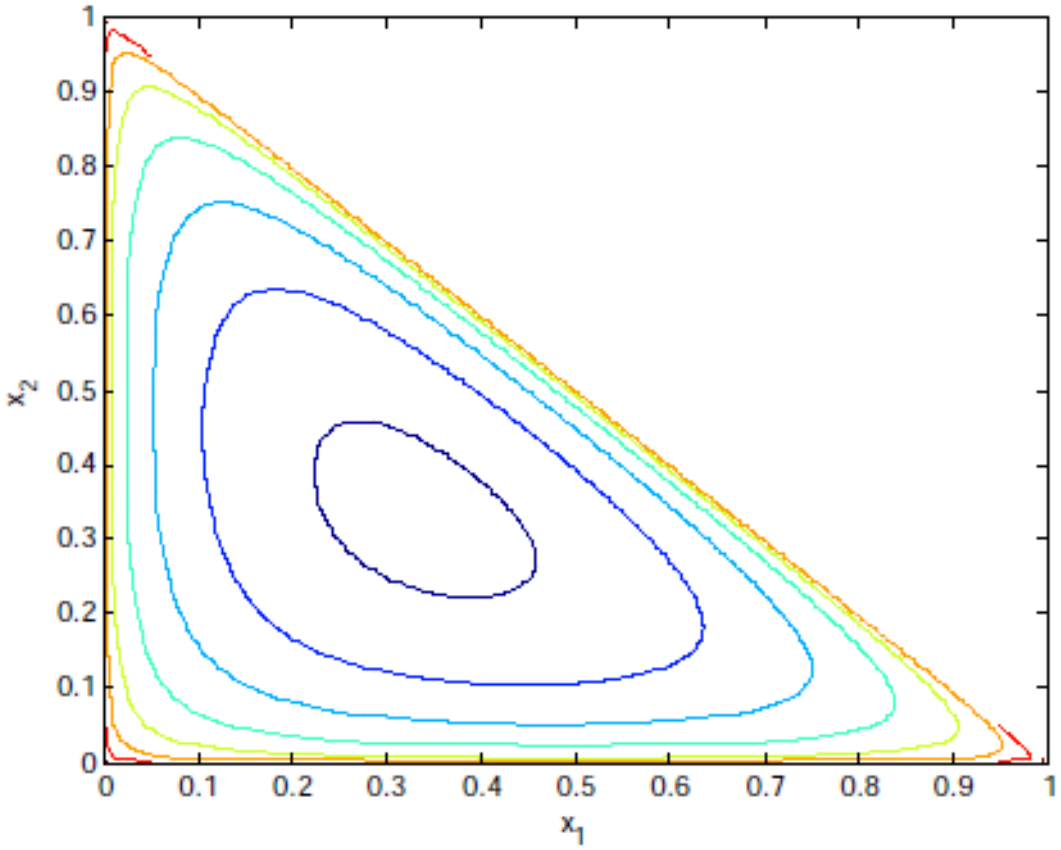
$$x^{(k+1)} := x^{(k)} - F(x^{(k)})^{-1} g^{(k)}.$$

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

- Can be viewed an iteration for solving $g(x) = 0$ using its Jacobian $F(x)$.

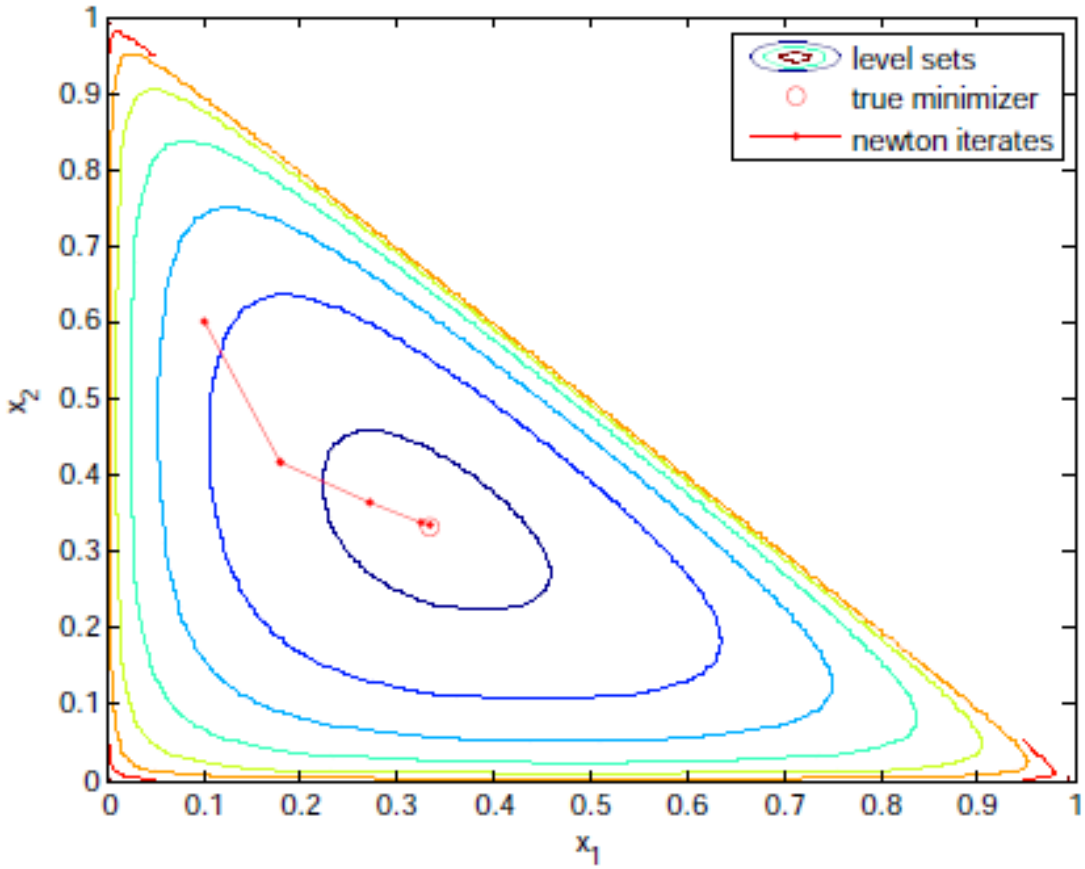
Example

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



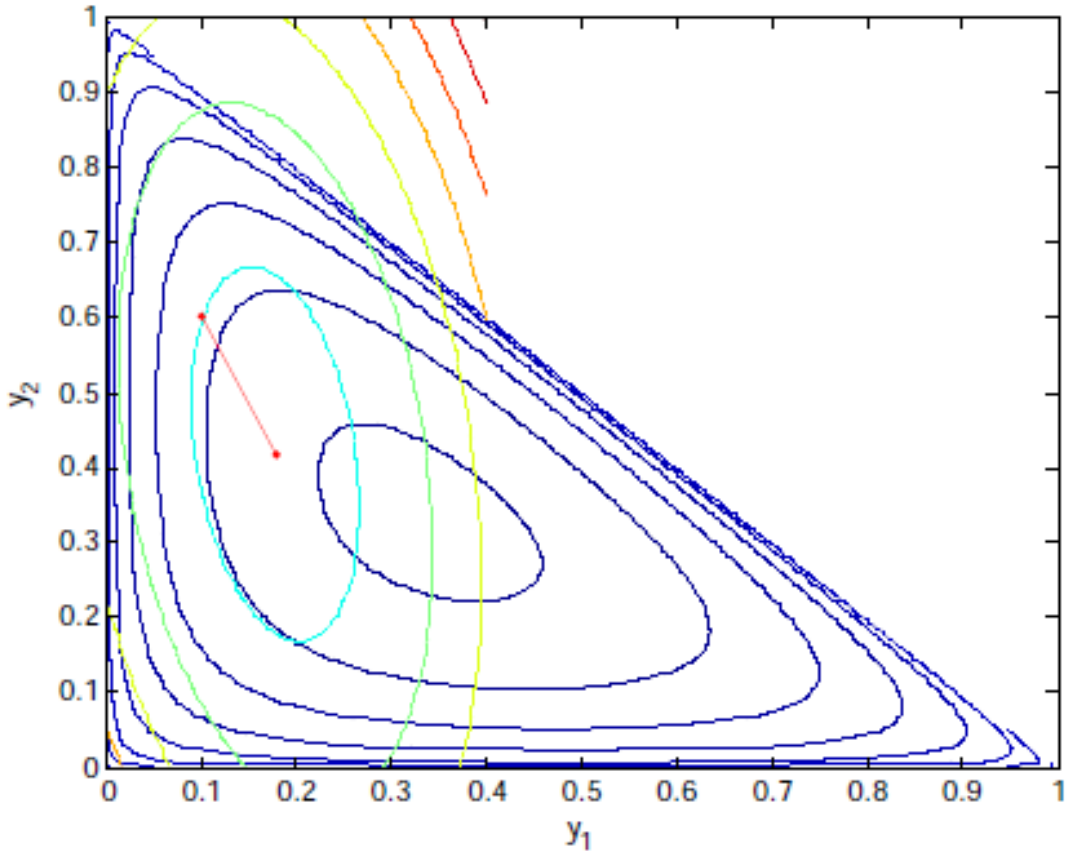
Example

Start Newton's method from $[\frac{1}{10}; \frac{6}{10}]$



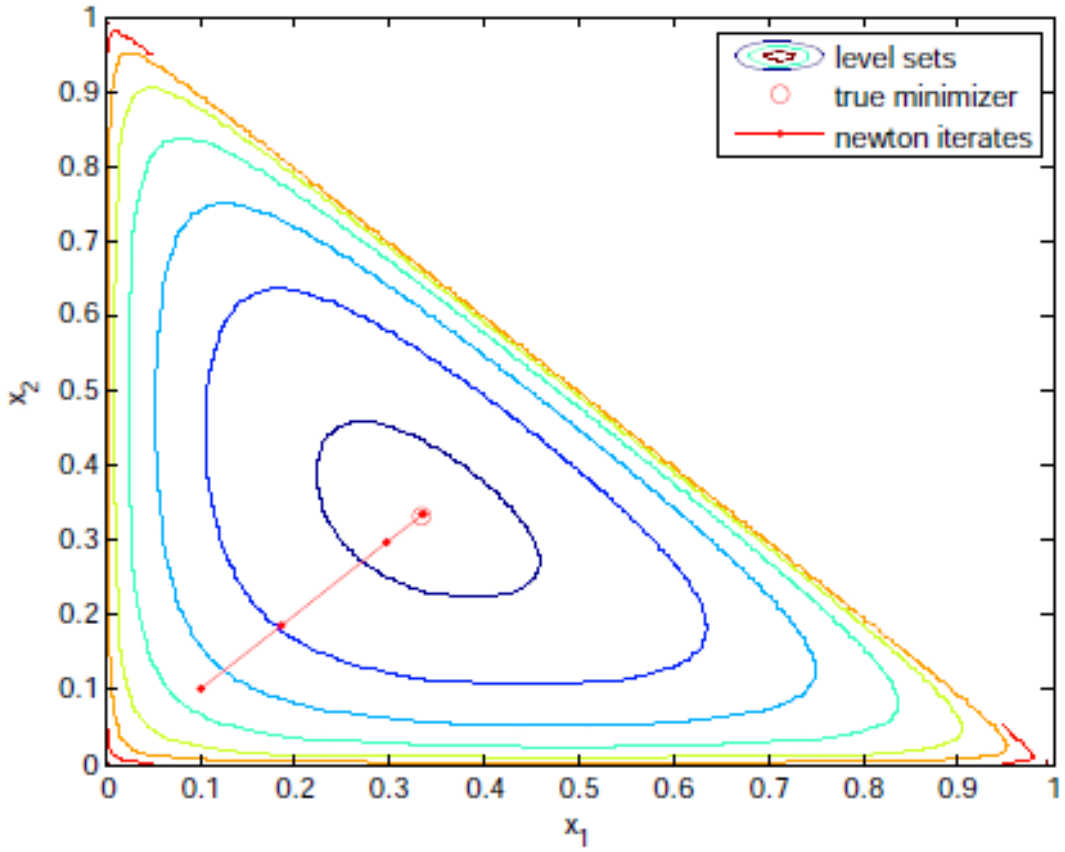
Example

$f(x_1, x_2)$ and its quadratic approximation $q(x_1, x_2)$ at $[\frac{1}{10}; \frac{6}{10}]$ share the same value, gradient and Hessian at $[\frac{1}{10}; \frac{6}{10}]$. The new point minimizes $q(x_1, x_2)$



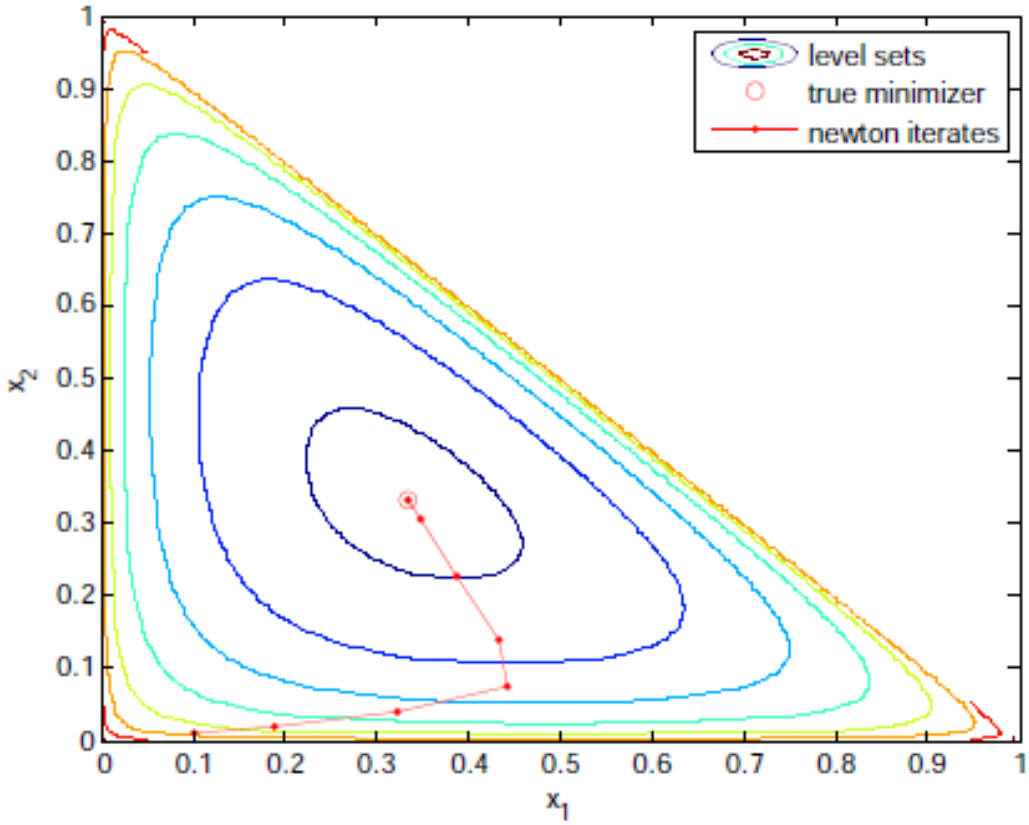
Example

Start Newton's method from $[\frac{1}{10}; \frac{1}{10}]$



Example

Start Newton's method from $[\frac{1}{10}; \frac{1}{100}]$



Aware of the drawbacks

Unlike the move along $-\nabla f(x^{(k)})$, where a sufficiently small step size guarantees the objective decrease, Newton's method jumps to a potentially distant point. This makes it vulnerable.

- recall in $1\mathrm{D}$, $f'' < 0$ can cause divergence
- even if $F(x^{(k)}) \succ 0$ (positive definite), objective descent is not guaranteed

Nonetheless, Newton's method has superior performance when starting near the solution.

Analysis: quadratic function minimization

- The objective function

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

  Assumption: Q is symmetric and invertible

$$g(x) = Qx - b$$
$$F(x) = Q.$$

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

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

  The solution is obtained in one step.

Analysis: how fast is Newton's method?

(assumption: Lipschitz continuous Hessian near x^*)

- Let $e^{(k)} := x^{(k)} - x^*$

Theorem

Suppose $f \in \mathcal{C}^2$, F is Lipschitz continuous near x^*, and $\nabla f(x^*) = 0$. If $x^{(k)}$ is sufficiently close to x^* and $F(x^*) \succ 0$, then $\exists C > 0$ such that

$$\|e^{(j+1)}\| \leq C\|e^{(j)}\|^2, j = k, k+1, \cdots$$

Just a sketch proof.

Since F is Lipschitz around x^* and $F(x^*) \succ 0$, we can have

- $(F(x) - cI) \succ 0$ for some $c > 0$ for all x in a small neighborhood of x^*.
- Thus, $\|F(x)^{-1}\| < c^{-1}$ for all x in the neighborhood.

$e^{(j+1)} = e^{(j)} + d^{(j)}$ where $d^{(j)} = -F(x^{(j)})^{-1}g^{(j)}$.

Taylor expansion near $x^{(j)}$: $0 = g(x^*) = g(x^{(j)}) - F(x^{(j)})e^{(j)} + O(\|e^{(j)}\|^2)$.

Thus $e^{(j+1)} = e^{(j)} + d^{(j)} = F(x^{(j)})^{-1}O(\|e^{(j)}\|^2) = O(\|e^{(j)}\|^2)$. Argue that $x^{(j+1)}, x^{(j+2)}, \cdots$ stay in the neighborhood. ■

Asymptotic rates of convergence

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

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

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

To distinguish superlinear rates of convergence, we check

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

- 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

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

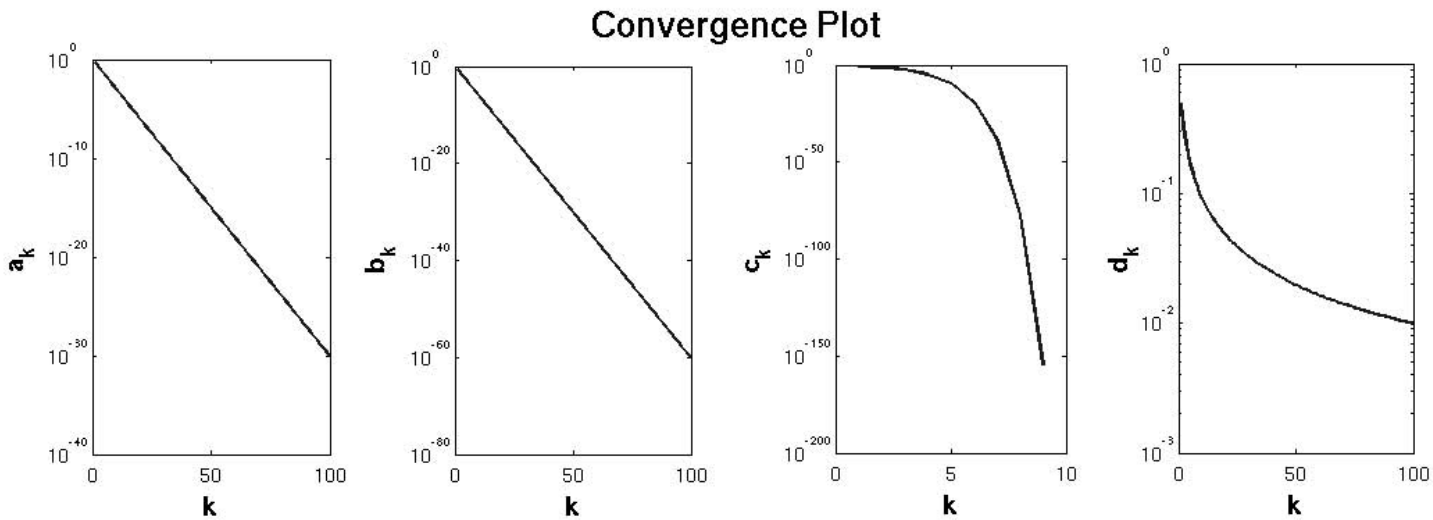


Figure: "semilogy" plots (wikipedia)

Another example

Let $C = 100$.

k	$(1/k^2)$	Ce^{-k}	$Ce^{-k^{1.618}}$	Ce^{-k^2}
1	1.0e0	3.7e2	3.7e2	3.7e2
3	3.3e-1	5.0e1	2.7e0	1.2e-1
5	2.0e-1	6.7e-0	1.3e-3	1.4e-8
7	1.4e-1	9.1e-1	7.6e-8	5.2e-19
9	1.1e-1	1.2e-1	6.4e-13	6.6e-33

Comments:

- the constant C is not important in superlinear convergence
- even with a big C, higher-order convergence will quickly catch up
- the constant C is more important in lower-order convergence
- superlinear convergence is shockingly fast!

Analysis: descent direction

Theorem

If the Hessian $F(x^{(k)}) \succ 0$ *(positive definite) and* $g^{(k)} = \nabla f(x^{(k)}) \neq 0$*, 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*

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

Proof. Let $\phi(\alpha) := f(x^{(k)} + \alpha d^{(k)})$. Then $\phi'(\alpha) := \nabla f(x^{(k)} + \alpha d^{(k)})^T d^{(k)}$. Since $F(x^{(k)}) \succ 0$ and $g^{(k)} \neq 0$, we have $F(x^{(k)})^{-1} \succ 0$

$$\phi'(0) := \nabla f(x^{(k)} + \alpha d^{(k)})^T d^{(k)} = -{g^{(k)}}^T F(x^{(k)})^{-1} g^{(k)} < 0.$$

Finally, apply first-order Taylor expansion to $\phi(\alpha)$ to get the result. ■

Two more issues with Newton's method

Hessian evaluation:

- When the dimension n is large, obtain $F(x^{(k)})$ can be computationally expensive
- We will study quasi-Newton methods to alleviate this difficulty (in a future lecture)

Indefinite Hessian:

- When the Hessian is not positive definite, the direction is not necessarily descending.
- There are simple modifications.

Modified Newton's method

Strategy:

- use $F(x^{(k)})$ if $F(x^{(k)}) \succ 0$ and $\lambda_{\min}(F(x^{(k)})) > \epsilon$; otherwise,
- use $\widehat{F}(x^{(k)}) = F(x^{(k)}) + E$ so that $\widehat{F}(x^{(k)}) \succ 0$ and $\lambda_{\min}(\widehat{F}(x^{(k)})) > \epsilon$.

Method 1 (Greenstadt): replace any tiny or negative eigenvalues by

$$\delta = \max\{\epsilon_{\text{machine}}, \epsilon_{\text{machine}} \|H\|_\infty\}$$

where $\|H\|_\infty = \max_{i=1,\cdots,n} \sum_{j=1}^{n} |h_{ij}|$. This is computationally expensive.

Method 2 (Levenberg-Marquardt): Was proposed for least-squares but works here. Replace

$$\widehat{F} \leftarrow F + \gamma I$$

It shifts every eigenvalue of F up by γ.

Modified Newton's method

Method 3 (advanced topic: modified Cholesky / Gill-Murray): Any symmetric matrix $A \succ 0$ can be factored as

$$A = \bar{L}\bar{L}^T \text{ or } A = LDL^T,$$

where L and $\bar{L}$ are lower triangular, D is positive diagonal, and L has ones on its main diagonal.

Properties of the Cholesky factorization:

- Very useful in solving linear systems of equations. Reduces a system to two backsolves.
- If $A \not\succ 0$ (indefinite but still symmetric), D has zero or negative element(s) on its diagonal.

Modified Newton's method

- The factorization is stable if A is positive definite. (Small errors in A will not cause large errors in L or D.) Example:

$$\begin{bmatrix} 1 & a \\ a & 1 \end{bmatrix} = \begin{bmatrix} 1 & \\ a & 1 \end{bmatrix} \begin{bmatrix} 1 & \\ & 1-a^2 \end{bmatrix} \begin{bmatrix} 1 & a \\ & 1 \end{bmatrix}$$

- If $A \leftarrow A + vv^T$, the factorization can be updated to a product form (avoiding the factorization from scratch, which is more expensive).
- If A is sparse, Cholesky with pivots keeps L sparse with moderately more zeros
- The cost is $n^3/6 + O(n^2)$, roughly half of Gaussian elimination.

Modified Newton's method

Forsgren, Gill, Murray: perform pivoted Cholesky factorization. That is, permute the matrix at each step to pull the largest remaining diagonal element to the pivot position

The effect: postpone the modification and keeps it as small as possible.

When no acceptable element remains

$$\begin{bmatrix} A_{11} & A_{12} \\ A_{21} & A_{22} \end{bmatrix} = \begin{bmatrix} L_{11} & \\ L_{21} & I \end{bmatrix} \begin{bmatrix} D_1 & \\ & D_2 \end{bmatrix} \begin{bmatrix} L_{11}^T & L_{21}^T \\ & I \end{bmatrix}$$

replace D_2 (not necessarily diagonal!) by a positive definition matrix and complete the factorization.

No extra work if the Cholesky factorization is taken in the outer-product form.

The Cholesky factorization also tells if the current point is a minimizer or a saddle point.

Modified Newton's method for saddle point

Suppose we are at a saddle point $\bar{x}$. Then $\bar{g} = \nabla f(\bar{x}) = 0$ and 2nd-order approximation

$$f(\bar{x}+d) \approx q(d) := f(\bar{x}) + \underbrace{\bar{g}^T d}_{=0} + \frac{1}{2} d^T F(\bar{x}) d.$$

How do we descend?

Greenstadt: pick $d = \sum_{i:\lambda_i<0} \alpha_i u_i$, where $\alpha_i > 0$ and (λ_i, u_i) are eigen-pairs of $F(x)$. d is a positive linear combination of the negative curvature directions. Then, $d^T F(\bar{x}) d < 0$ and $q(d) < f(\bar{x})$.

Cholesky method: recall D_2 correspond to the negative curvatures. If entry d_{ij} of D_2 has the largest absolute value among all entries of D_2, pick

$$L^T d = e_i - \text{sign}(d_{ij}) e_j.$$

Then, $d^T F(x) d = d^T L D L^T d < 0$.

Overview of the Gauss-Newton method

- A modification to Newton's method, solves nonlinear least squares, very popular
- Pros: second derivatives are no longer computed
- Cons: does not apply to general problems
- Can be improved by line search, Levenberg-Marquardt, etc.

Nonlinear least squares

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

$$\min_x \ \frac{1}{2} \sum_{i=1}^{m} (r_i(x))^2.$$

- Define $r = [r_1, \cdots, r_m]^T$. Then we have

$$\min_x \ f(x) = \frac{1}{2} r(x)^T r(x).$$

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

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

- Define the Jacobian of r

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

Then, we have

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

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

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

- Let $S(x)$ be formed by (k,j)th components

$$r_i(x)\frac{\partial^2 r_i}{\partial x_k \partial x_j}(x)$$

- Then, we have $F(x) = J(x)^T J(x) + S(x)$
- Therefore, Newton's method has the iteration

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

The Gauss-Newton method

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

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

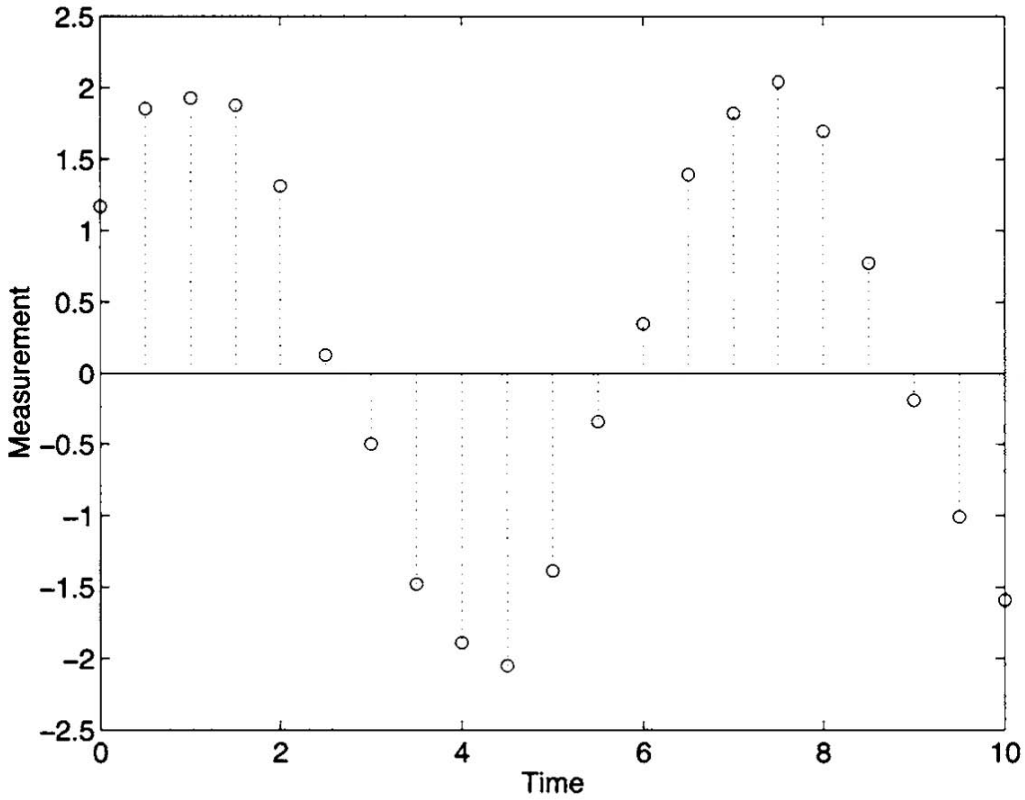
- A potential problem is that $J(x)^T J(x) \not\succ 0$ and $f(x^{(k+1)}) \geq f(x^{(k)})$. Fixes: line search, Levenberg-Marquardt, and Cholesky/Gill-Murray.

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: $(t_i, y_i), i = 1, \cdots, 21$.



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

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

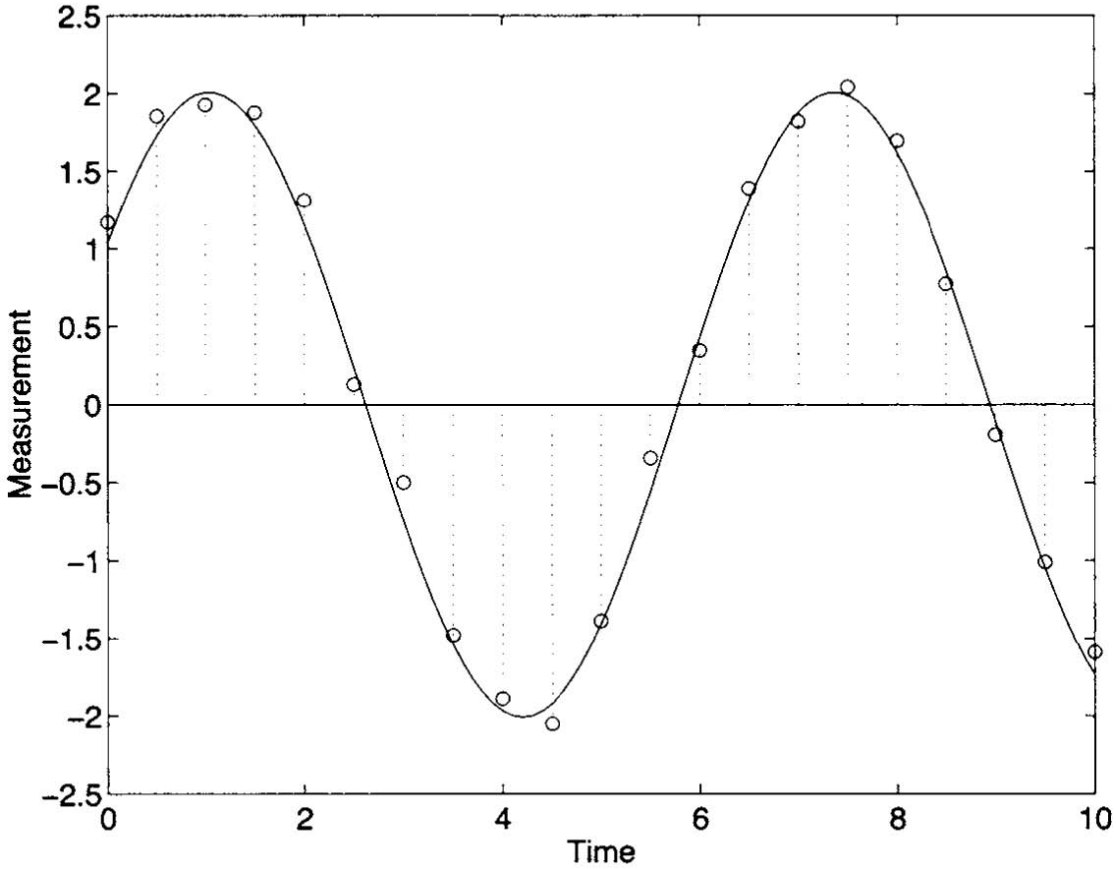
- Problem

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

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

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

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


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

Conclusions

Although Newton's method has many issues, such as

- the direction can be ascending if $F(x^{(k)}) \not\succ 0$
- 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 Q
- second-order convergence rate near the solution if F is Lipschtiz
- a number of modifications that address the issues.